ON TWO-SAMPLE DATA ANALYSIS BY EXPONENTIAL MODEL

A Dissertation

by

SUJUNG CHOI

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

August 2005

Major Subject: Statistics

ON TWO-SAMPLE DATA ANALYSIS BY EXPONENTIAL MODEL

A Dissertation

by

SUJUNG CHOI

Submitted to Texas A&M University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Approved by:

Chair of Committee,	Emanuel Parzen
Committee Members,	Bani Mallick
	Michael Sherman
	David R. Larson
Head of Department,	Simon Sheather

August 2005

Major Subject: Statistics

ABSTRACT

On Two-Sample Data Analysis By Exponential Model. (August 2005)

Sujung Choi, B.A., Yonsei University;

M.A., Yonsei University

Chair of Advisory Committee: Dr.Emanuel Parzen

We discuss two-sample problems and the implementation of a new two-sample data analysis procedure. The proposed procedure is based on the concepts of mid-distribution, design of score functions, components, comparison distribution, comparison density and exponential model. Assume that we have a random sample $X_1, \cdots, X_m$ from a continuous distribution $F(y) = P(X_i \leq y), i = 1, \cdots, m$ and a random sample $Y_1, \cdots, Y_n$ from a continuous distribution $G(y) = P(Y_i \leq y), i = 1, \cdots, n$. Also assume independence of the two samples. The two-sample problem tests homogeneity of two samples and formally can be stated as $H_0 : F = G$. To solve the two-sample problem, a number of tests have been proposed by statisticians in various contexts. Two typical tests are the two-sample $t-$test and the Wilcoxon's rank sum test. However, since they are testing differences in locations, they do not extract more information from the data as well as a test of the homogeneity of the distribution functions. Even though the Kolmogorov-Smirnov test statistic or Anderson-Darling tests can be used for the test of $H_0 : F = G$, those statistics give no indication of the actual relation of F to G when $H_0 : F = G$ is rejected. Our goal is to learn why it was rejected. Our approach gives an answer using graphical tools which is a main property of our approach. Our approach is functional in the sense that the parameters to be estimated are probability density functions. Compared with other statistical tools for two-sample problems such as the t-test or the Wilcoxon rank-sum test, density esti-

mation makes us understand the data more fully, which is essential in data analysis. Our approach to density estimation works with small sample sizes, too. Also our methodology makes almost no assumptions on two continuous distributions F and G. In that sense, our approach is nonparametric. Our approach gives graphical elements in two-sample problem where exist not many graphical elements typically. Furthermore, our procedure will help researchers to make a conclusion as to why two populations are different when H_0 is rejected and to give an explanation to describe the relation between F and G in a graphical way.

To everyone who has been praying for me

ACKNOWLEDGEMENTS

Graduate study in the Department of Statistics at Texas A&M means a lot to me. I have spent a fifth of my life here and learned so many precious things about statistics and life. In this respect, I am indebted to many of the faculty of the Department of Statistics, not only for discussions related to this dissertation but also for their time and effort while teaching the several courses I attended. In this regard, I would like to thank Professor Bani Mallick, Professor Michael Sherman and Professor David Larson from whom I learned much.

My advisor, Professor Manny Parzen is a great scholar and philosopher. It was a great opportunity for me to work with him. From him, I learned philosophy and attitude on research as well as knowledge of statistics. Also, he trained me to be a statistician with his great vitality and intelligence. He spared no pains to give all the detailed comments on my dissertation which greatly improved my dissertation. I really appreciate what he has done for me. Also, I want to give my special thanks to Professor Michael Longnecker and Marilyn Randall who have been taking care of all necessary documents to maintain my student status. Without their help, it would have been difficult to concentrate on my study in the states.

This work would never have been done without understanding, love and support of my family in Korea. In any situations, they supported my study and believed in me. Since I know that it was not always easy for them, I would like to express my appreciation from the bottom of my heart. Also, there are friends who helped me in several ways here, in College Station. I would like to express my special thanks to them since they made me feel relaxed whenever I was under great stress. After all, I could not have completed this work without other people's help.

TABLE OF CONTENTS

CHAPTER Page

ABSTRACT iii

ACKNOWLEDGEMENTS vi

TABLE OF CONTENTS vii

LIST OF TABLES ix

LIST OF FIGURES x

CHAPTER

I INTRODUCTION 1

1.1. The Two-Sample Problem 1
1.2. Outline of This Dissertation 3

II COMPARISON DISTRIBUTION FUNCTION AND COMPARISON DENSITY FUNCTION 4

2.1. Introduction 4
2.2. Population Concepts 4
2.3. Sample Comparison Distribution and Comparison Density Functions 12

III EXPONENTIAL MODEL WITH COMPONENTS 22

3.1. Introduction 22
3.2. Basic Concepts 24
3.2.1. Mid-distribution Functions 24
3.2.2. Design of Score Functions 26
3.2.3. Components 27
3.2.4. Exponential Model Approach and Comparison Density Estimation 31
3.2.4.1. Maximum Entropy Interpretation of Exponential Model Approach 33
3.2.4.2. Estimation of Coefficients of An Exponential Model 36

IV TWO-SAMPLE DATA ANALYSIS PROCEDURE 41

CHAPTER Page

4.1. Introduction 41
4.2. Algorithm 41
4.3. Summary and Discussion: Stress Data 43

V EXAMPLES AND APPLICATIONS 49

5.1. Introduction 49
5.2. Radon Cancer Data 49
5.3. Explanatory Data Analysis 51
5.4. Two-sample Data Analysis Using Exponential Model Approach 51
5.5. Summary and Discussion: Radon Cancer Data 53
5.6. Simulation Results 63
5.6.1. Case 1: Same Distributions, Same Locations, and Same Scales 65
5.6.2. Case 2: Same Locations, Scales but Different Distributions 65
5.6.3. Case 3: Same Locations, Different Scales and Same Distributions 65
5.6.4. Case 4: Different Locations, but Same Scales and Distributions 65
5.6.5. Case 5: Same Locations, but Different Scales and Distributions 65
5.6.6. Case 6: Different Locations, Same Scales and Different Distributions 66
5.6.7. Case 7: Different Locations, Scales but Same Distributions 66
5.6.8. Case 8: Different Locations, Scales and Distributions 66
5.6.9. Summary and Discussion 66

VI CONCLUSION 89

6.1. Concluding Remarks 89
6.2. Problems for Future Study 90

REFERENCES 91

APPENDIX 95

VITA 99

LIST OF TABLES

TABLE		Page
1	Diastolic blood pressure (mmHg)	13
2	Sample distribution function for control group : $F_m(x) = \sum_{t=1}^{m} I(X_t \leq x)/m$, $m = 11$	14
3	Sample distribution function for stress group : $G_n(y) = \sum_{t=1}^{n} I(Y_t \leq y)/n$, $n = 11$. The number in () means the number of occurrences of the corresponding observation.	14
4	Sample pooled distribution function $H_N(z) = \sum_{t=1}^{m} I(X_t \leq z)/N + \sum_{t=1}^{n} I(Y_t \leq z)/N$, $N = 22$	15
5	Sample mid-distribution function $H_N^{mid}(z)$: $H_N^{mid}(z) = H_N(z) - .5p_H^{\sim}(z)$	39
6	Inner product of score functions to verify orthonormality with stress data	40
7	$\theta_j^{\wedge}$ values up to order 3 through Newton-Raphson iteration with stress data	40
8	Score function value up to order 4 with stress data	45
9	Radon concentration levels	50
10	Sample pooled distribution function H_N. The number in () means the number of occurrences of the corresponding observation.	54
11	Inner product of score functions to verify orthonormality with radon cancer data	57
12	$\theta_j^{\wedge}$ values up to order 2 through Newton-Raphson iteration with radon cancer data	59
13	All possible cases according to differences in either locations or scales or distributions. "0" means that there are no differences between two samples and "1" means there are differences between two samples	64

LIST OF FIGURES

FIGURE Page

1 Plots of unpooled comparison density function of normal distributions with difference in locations assuming $F = N(0,1)$ and $G = N(\theta,1)$. Unpooled comparison density is $\log d(u;F,G) = \big(\theta\Phi^{-1}(u) - .5\theta^2\big)$ where $\Phi^{-1}(u)$ is the inverse function of $F = N(0,1)$. As the difference in locations is getting greater, comparison density is getting farther from $\log d(u) = 0 (d(u) = 1)$. Also, $\log d(u)$ is monotone. 9

2 Plots of unpooled comparison density function of normal distributions with difference in scales assuming $F = N(0,1)$ and $G = N(0,\theta^{-2})$. Unpooled comparison density is $\log d(u;F,G) = \log\theta - .5\big(\Phi^{-1}(u)\big)^2(\theta^2-1)$ where $\Phi^{-1}(u)$ is the inverse function of $F = N(0,1)$. As the difference in scales is getting greater, comparison density is getting farther from $d(u) = 1 (\log d(u) = 0)$. Also, $\log d(u)$ is quadratic and symmetric. 10

3 Plot of sample unpooled comparison distribution function $D^{\sim}(u;F,G)$ with stress data. In this case, two properties of comparison distribution function$(D^{\sim}(0;G,F) = 0$ and ,$D^{\sim}(1;G,F) = 1)$ are not satisfied. 17

4 Plot of sample unpooled comparison distribution function $D^{\sim}(u;G,F)$ with stress data. 18

5 Plot of sample pooled comparison distribution function $D^{\sim}(u;H,F)$ with stress data. 19

6 Plot of sample pooled comparison density function $d^{\sim}(u;H,F)$ with stress data. 20

7 Side by side boxplot with stress data. 21

8 Sample mid-distribution score functions up to order 4 using stress data. $\psi_j\big(H_N^{-1}\big(u\big)\big)$ for $u_{i-1} < u < u_i$ and $H_N^{-1}(u_i) = z_i$. 34

FIGURE Page

9 $d^{\wedge}(u; H, F)$: Estimated comparison density function through exponential model approach with stress data. The step function, the quartile density is added to the graph to see how exponential model approach works. The quartile density is defined for $i = 1, 2, 3, 4$ by $dQ_k(u) = 4\{D^{\sim}\big(i(.25)\big) - D^{\sim}\big((i-1)(.25)\big)\}$, $(i-1).25 < u < i(.25)$. For lower value of blood pressure($u < 0.25$), the comparison density is greater than 1, indicating a great frequency of observations in control group. 46

10 95% bootstrap confidence interval of $d^{\wedge}(u; H, F)$: For better interpretation, bootstrap confidence interval is added and the confidence interval is computed through percentile method with 500 bootstrap samples. Since the confidence interval does not include uniform density $d_0(u)$, we conclude that the distributions of the blood pressure level of two groups are different and stress does have an effect on blood pressure level. 47

11 $D^{\wedge}(u; H, F)$: Estimated comparison distribution function with stress data. Since estimated comparison distribution function goes with $D^{\sim}(u; H, F)$ very well, we conclude that our exponential model estimation is working properly. 48

12 Sample pooled comparison distribution function with radon cancer data. 55

13 Sample mid-distribution score functions up to order 4 using radon cancer data. $\psi_j\big(H_N^{-1}\big(u\big)\big)$ for $u_{i-1} < u < u_i$ and $H_N^{-1}(u_i) = z_i$. 56

14 $d^{\sim}(u; H.F)$: Sample comparison density function with radon cancer data. 58

15 $d^{\wedge}(u; H, F)$: Estimated comparison density function through exponential model approach with two components with radon cancer data. 60

16 $D^{\wedge}(u; H, F)$: Estimated comparison distribution function with radon cancer data with 2 components. Since estimated comparison distribution function goes with $D^{\sim}(u; H, F)$ very well, we conclude that our exponential model estimation is working properly. 61

FIGURE Page

17 95% bootstrap confidence interval of $d^{\wedge}(u; H, F)$: For better interpretation, bootstrap confidence interval is added and the confidence interval is computed through percentile method with 500 bootstrap samples. Since the confidence interval does not include uniform density $d_0(u)$, we conclude that the distributions of the radon concentration level of two groups are different and radon does have an effect on childhood cancer incidence. 62

18 Case 2: Same locations, scales but different distributions: Probability density functions of $X \sim Normal(1, 1^2)$ and $Y \sim Gamma(1, 1)$. 68

19 Case 2: Same locations, scales but different distributions: $d^{\wedge}(u; H, F)$: Estimated comparison density function with $X \sim Normal(1, 1^2)$ and $Y \sim Gamma(1, 1)$. 2nd and 3rd order score functions were selected($\mathcal{C} = \{2, 3\}$). 69

20 Case 2: Same locations, scales but different distributions: $D^{\wedge}(u; H, F)$: Estimated comparison distribution function with $X \sim Normal(1, 1^2)$ and $Y \sim Gamma(1, 1)$. 70

21 Case 3: Same locations, different scales and same distributions: Probability density functions of $X \sim Normal(0, 5^2)$ and $Y \sim Normal(0, 1^2)$. 71

22 Case 3: Same locations, different scales and same distributions: $d^{\wedge}(u; H, F)$: Estimated comparison density function with $X \sim Normal(0, 5^2)$ and $Y \sim Normal(0, 1^2)$. Only 2nd order score function was selected($\mathcal{C} = \{2\}$). 72

23 Case 3: Same locations, different scales and same distributions: $D^{\wedge}(u; H, F)$: Estimated comparison distribution function with $X \sim Normal(0, 5^2)$ and $Y \sim Normal(0, 1^2)$ 73

24 Case 4: Different locations, but same scales and distributions: Probability density functions of $X \sim Normal(0, 1^2)$ and $Y \sim Normal(3, 1^2)$. 74

FIGURE Page

25 Case 4: Different locations, but same scales and distributions: $d^{\wedge}(u; H, F)$: Estimated comparison density function with $X \sim Normal(0, 1^2)$ and $Y \sim Normal(3, 1^2)$. Only 1st order component was selected($\mathcal{C} = \{1\}$). 75

26 Case 4: Different locations, but same scales and distributions: $D^{\wedge}(u; H, F)$: Estimated comparison distribution function with $X \sim Normal(0, 1^2)$ and $Y \sim Normal(3, 1^2)$ 76

27 Case 5: Same locations, but different scales and distributions: Probability density functions of $X \sim Normal(2, 1^2)$ and $Y \sim Gamma(1, 2)$. 77

28 Case 5: Same locations, but different scales and distributions: $d^{\wedge}(u; H, F)$: Estimated comparison density function with $X \sim Normal(2, 1^2)$ and $Y \sim Gamma(1, 2)$. 2nd, 3rd, and 4th order score functions were selected($\mathcal{C} = \{2, 3, 4\}$). 78

29 Case 5: Same locations, but different scales and distributions: $D^{\wedge}(u; H, F)$: Estimated comparison distribution function with $X \sim Normal(0, 1^2)$ and $Y \sim Normal(3, 1^2)$ 79

30 Case 6: Different locations, same scales and different distributions: Probability density functions of $X \sim Normal(0, \sqrt{2}^2)$ and $Y \sim Gamma(1, 2)$. 80

31 Case 6: Different locations, same scales and different distributions: $d^{\wedge}(u; H, F)$: Estimated comparison density function with $X \sim Normal(0, \sqrt{2}^2)$ and $Y \sim Gamma(1, 2)$. 1st and 2nd order score functions were selected($\mathcal{C} = \{1, 2\}$). 81

32 Case 6: Different locations, same scales and different distributions: $D^{\wedge}(u; H, F)$: Estimated comparison distribution function with $X \sim Normal(0, \sqrt{2}^2)$ and $Y \sim Gamma(1, 2)$ 82

33 Case 7: Different locations, scales but same distributions: Probability density functions of $X \sim Normal(0, 1^2)$ and $Y \sim Normal(3, 2^2)$. 83

FIGURE Page

34 Case 7: Different locations, scales but same distributions: $d^{\wedge}(u; H, F)$: Estimated comparison density function with $X \sim Normal(0, 1^2)$ and $Y \sim Normal(3, 2^2)$. 1st and 3rd order score functions were selected($\mathcal{C} = \{1, 3\}$). 84

35 Case 7: Different locations, scales but same distributions: $D^{\wedge}(u; H, F)$: Estimated comparison distribution function with $X \sim Normal(0, 1^2)$ and $Y \sim Normal(3, 2^2)$ 85

36 Case 8: Different locations, scales and distributions: Probability density functions of $X \sim Normal(0, 1^2)$ and $Y \sim Gamma(2/3, 2)$. . 86

37 Case 8: Different locations, scales and distributions: $d^{\wedge}(u; H, F)$: Estimated comparison density function with $X \sim Normal(0, 1^2)$ and $Y \sim Gamma(2/3, 2)$. 1st and 3rd order score functions were selected($\mathcal{C} = \{1, 3\}$). 87

38 Case 8: Different locations, scales and distributions: $D^{\wedge}(u; H, F)$: Estimated comparison distribution function with $X \sim Normal(0, 5)$ and $Y \sim Gamma(2/3, 2)$ 88

CHAPTER I

INTRODUCTION

1.1. The Two-Sample Problem

Assume that we have a random sample $X_1, \cdots, X_m$ from a continuous distribution $F(y) = P(X_i \leq y), i = 1, \cdots, m$ and a random sample $Y_1, \cdots, Y_n$ from a continuous distribution $G(y) = P(Y_i \leq y), i = 1, \cdots, n$. Also assume independence of the two samples. The two-sample problem is about homogeneity of the two samples and formally can be stated as $H_0 : F = G$. Even though we stated the two-sample problem in terms of distributions, the two-sample problem could be homogeneity in locations or in scales of the two samples. Borovkov (1998) provides several examples of the two-sample problem.

- A comparison of two processing techniques on the crops of some variety cereals.
- A test of the effect of a new drug by means of comparing the state of patients in two groups, one taking the drug and the other(the control group) not.
- A comparison of the car accident ratios in two cities.

To solve the two-sample problem, a number of tests were proposed by statisticians in various contexts. Some of the tests need specific assumptions on the nature of two distributions. According to assumptions on distributions, we classify tests into

---

The format and style of this dissertation follow that of *Biometrics* .

parametric test and nonparametric tests. Two typical tests are the two-sample t-test and the Wilcoxon's rank sum test respectively. However, since they are testing the differences in locations, they do not extract more information from the data as well as a test of homogeneity of distribution functions. Even though the Kolmogorov-Smirnov test statistic or the Anderson-Darling test can be used for test of $H_0 : F = G$, those statistics give no indication of the actual relation of F to G even though $H_0 : F = G$ is rejected. The point is why it was rejected. But most two-sample techniques can not answer this point. Our approach gives an answer using graphical tools which is a main property of our approach. Our approach is unified in the sense that graphs and tests are derived from a common foundation, comparison distribution and comparison density. The comparison density is graphical in nature and carries information regarding the relation of f to g.

The goal of this dissertation is to discuss the two-sample problem and our main contribution will be to implement and illustrate a two-sample data analysis procedure which extracts more information from the data by a methodology that makes almost no assumptions on two continuous distributions F and G. In that sense, our approach is nonparametric. Also, our approach gives graphical elements in the two-sample problem where typically exist not many graphical elements such as side by side boxplot, Q-Q plots and histograms which are not very informative. Also, our procedure will help researchers to make a conclusion to why two populations are different when H_0 is rejected and to give an explanation to describe the relation between F and G in a graphical way.

1.2. Outline of This Dissertation

This dissertation is composed of six chapters and an appendix. Chapter I is an introduction of the two-sample problem. In Chapter II, concepts of comparison distribution function and comparison density function are discussed. Especially in section 2.3, sample versions of those functions are discussed for implementation in practice while population concepts are provided in section 2.2. Also, properties of comparison distribution and density functions are reviewed. In section 2.3, a real data set is used to illustrate sample concepts of those functions.

Chapter III examines exponential model with components with other necessary concepts such as mid-distribution functions, score functions and components. In section 3.2.4, we discuss exponential model approach to comparison density function. Maximum entropy interpretation of exponential model is also given in section 3.2.4.1. And the following section 3.2.4.2 is dedicated on estimation of coefficient of exponential model.

Chapter IV provides an algorithm to solve the two-sample problem. That algorithm is based on comparison density estimation through exponential model approach. Chapter V applies the algorithm of Chapter IV to a real data set and to simulated data sets. Chapter VI presents conclusions and future research interests.

Appendix A gives some proofs of the theorems and properties stated in the previous chapters.

CHAPTER II

COMPARISON DISTRIBUTION FUNCTION AND COMPARISON DENSITY FUNCTION

2.1. Introduction

In this chapter, comparison distribution and comparison density functions are defined under the two-sample frame. As a graphical and functional type of test, Parzen (1983) introduced the concept of comparison density. In one-sample case, comparison distribution is comparing a model for a true distribution and the sample distribution. In the following sections, the comparison distribution and comparison density are defined and some properties of them are discussed.

2.2. Population Concepts

We can formulate the two-sample problem as the comparison of two continuous distribution functions F and G of variables X and Y respectively. Assume that we have a sample $X_1, \cdots, X_m$ from a continuous distribution F and a sample $Y_1, \cdots, Y_n$ from a continuous distribution G. Assume F and G have continuous densities f and g. Let $N = m + n$ and $\lambda_N = m/N$. To compare two continuous distributions F and G, we define two versions of comparison distributions, unpooled comparison distribution $D(u; F, G)$, and pooled comparison distribution function $D(u; H, F)$ where $0 \leq u \leq 1$ and H is defined by

$$H(y) = \lambda F(y) + (1 - \lambda)G(y) \tag{2.1}$$

assuming that $\lim_{N\to\infty} m/N = \lambda$ with $0 < \lambda < 1$. Define the following inverse functions at $0 \leq u \leq 1$:

$$\begin{aligned} F^{-1}(u) &= \inf\{y; F(y) \geq u\}, \\ G^{-1}(u) &= \inf\{y; G(y) \geq u\}, \\ H^{-1}(u) &= \inf\{y; H(y) \geq u\}. \end{aligned} \tag{2.2}$$

The unpooled comparison distribution function is defined as

$$D(u; F, G) = G(F^{-1}(u)), \quad 0 \leq u \leq 1 \tag{2.3}$$

with assumptions that $D(0; F, G) = 0$ and $D(1; F, G) = 1$. The pooled comparison distribution function is

$$D(u) = D(u; H, F) = F(H^{-1}(u)), \quad 0 \leq u \leq 1. \tag{2.4}$$

Research on comparing the two distributions has tended to focus on estimating the unpooled estimator. However, if F and G do not have the same support, the comparison distribution is not always rigorously definable. For example, suppose F is a distribution of incomes of men and G is a distribution of incomes of women. Then, the support of F may not be contained in that of G. Therefore, Parzen (1997) recommends to use the pooled comparison distribution. The properties of $D(u)$ are as follows:

- $D(0) = 0$.
- $D(1) = 1$.
- $D(u)$ is non-decreasing on $[0, 1]$.
- $D(u)$ is absolutely continuous on $[0, 1]$

Another problem of the unpooled comparison distribution is that the first two properties ($D(0) = 0$ and $D(1) = 1$) may not be satisfied with the sample unpooled comparison distribution which will be defined in the section 2.3. We will explain this in detail with an example in the section 2.3.

Derivatives of $D(u)$ are called the comparison density functions. The unpooled comparison density function $d(u; F, G)$ and the pooled comparison density function $d(u; H, F)$ are defined respectively by

$$\begin{aligned} d(u; F, G) &= D'(u; F, G) = \frac{g(F^{-1}(u))}{f(F^{-1}(u))}, \\ d(u; H, F) &= D'(u; H, F) = \frac{f(H^{-1}(u))}{h(H^{-1}(u))} \\ &= \frac{f(H^{-1}(u))}{\lambda f(H^{-1}(u)) + (1-\lambda) g(H^{-1}(u))}. \end{aligned} \tag{2.5}$$

We require $f(x) = 0$ implies $g(x) = 0$ in order for $d(u; F, G)$ to be well defined and to integrate to 1. Given a plot of $d(u)$, one can interpret the various shapes as indicating that the difference between two distributions(F and G for unpooled case and H and F for pooled case) is a difference in location or a difference in scale by the following known theorem.

THEOREM 2.1. *(Parzen (1998)) Assume $F = N(\theta_0, 1)$ and $G = N(\theta, 1)$; that is, the difference between two Normal distributions is due to a difference in location. Then, the unpooled comparison density satisfies*

$$\log\, d(u; F, G) = (\theta - \theta_0)\Phi^{-1}(u) - .5(\theta - \theta_0)^2. \tag{2.6}$$

where $F = \Phi$ which is the standard normal distribution. When $F = N(0, \theta_0^{-2})$ and $G = N(0, \theta^{-2})$, that is, if there is a difference in scale, unpooled comparison density satisfies

$$\log\, d(u; F, G) = \log\, \frac{\theta}{\theta_0} - .5\big(\Phi^{-1}(u)\big)^2(\theta^2 - \theta_0^2). \tag{2.7}$$

Proof

$$\begin{aligned} d(u;F,G) &= \frac{g\big(F^{-1}(u)\big)}{f\big(F^{-1}(u)\big)} \\ &= \frac{\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(\Phi^{-1}(u)-\theta)^2}{2}\right)}{\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{(\Phi^{-1}(u)-\theta_0)^2}{2}\right)} \\ \Rightarrow \quad \log\, d(u;F,G) &= (\theta-\theta_0)\Phi^{-1}(u) - .5(\theta^2-\theta_0^2). \end{aligned} \tag{2.8}$$

$$\begin{aligned} d(u;F,G) &= \frac{g\big(F^{-1}(u)\big)}{f\big(F^{-1}(u)\big)} \\ &= \frac{\frac{\theta}{\sqrt{2\pi}}\exp\left(-\frac{\theta^2\Phi^{-1}(u)^2}{2}\right)}{\frac{\theta_0}{\sqrt{2\pi}}\exp\left(-\frac{\theta_0^2\Phi^{-1}(u)^2}{2}\right)} \\ \Rightarrow \quad \log\, d(u;\theta) &= \log\,\frac{\theta}{\theta_0} - .5(\Phi^{-1}(u))^2(\theta^2-\theta_0^2). \end{aligned} \tag{2.9}$$

For pooled comparison density with difference in locations, assume $F = N(0,1)$ and $G = N(\theta,1)$. Then for a given λ, pooled distribution $H(y)$ is a mixture normal distribution.

$$H(y) = \lambda N(0,1) + (1-\lambda)N(\theta,1).$$

For pooled comparison density with difference in scales, assume $F = N(0,1)$ and $G = N(0,\theta^{-2})$. Then for a given λ, pooled distribution $H(y)$ is a mixture normal distribution.

$$H(y) = \lambda N(0,1) + (1-\lambda)N(0,\theta^{-2}).$$

In the case of pooled comparison density, we do not have a closed form like unpooled comparison density functions. Thus, pooled comparison density can be computed by simulations. In simulation, a very large sample of random variables from known

distributions $F(y)$ and $G(y)$ will be generated. Then, sample pooled comparison distribution and density function can be computed.

For the unpooled case, Figure 1 and Figure 2 present $\log d(u; F, G)$ for a variety of F and G. If two distributions F and G are homogeneous, $d(u; F, G) = 1$(or $\log d(u; F, G) = 0$). As the difference in locations is getting greater, comparison density is getting farther from $d(u; F, G) = 1$(or $\log d(u; F, G) = 0$). Also, $\log d(u; F, G)$ is monotone for location difference and quadratic for scale difference. As the difference in scales is getting greater, comparison density is getting farther from $d(u; F, G) = 1$(or $\log d(u; F, G) = 0$). Also, $\log d(u; F, G)$ is quadratic and symmetric since there is no difference in location.

Parzen (1983) gives some properties of pooled comparison density $d(u) = d(u; H, F)$.

- $0 \leq d(u) \leq 1/\lambda$
- $d(u) \to 0$ if $f \to 0$
- $d(u) \to 1/\lambda$ if $g \to 0$

For the proofs, see appendix A. From the definition of the pooled comparison density function, we can see the relationship between $d(u)$ and likelihood ratio(g/f). Parzen (1983) noted that

$$\frac{1}{d(u)} = \lambda + (1 - \lambda)\frac{g(H^{-1}(u))}{f(H^{-1}(u))} \tag{2.10}$$

which is derived from equation(2.5). If an estimate of g/f is not really desired, it is enough to know that $d(u) > 1$ if and only if $g(H^{-1}(u)) > f(H^{-1}(u))$. Also, even though g/f is not bounded, $d(u)$ is bounded between 0 and $1/\lambda$. Since the estimation

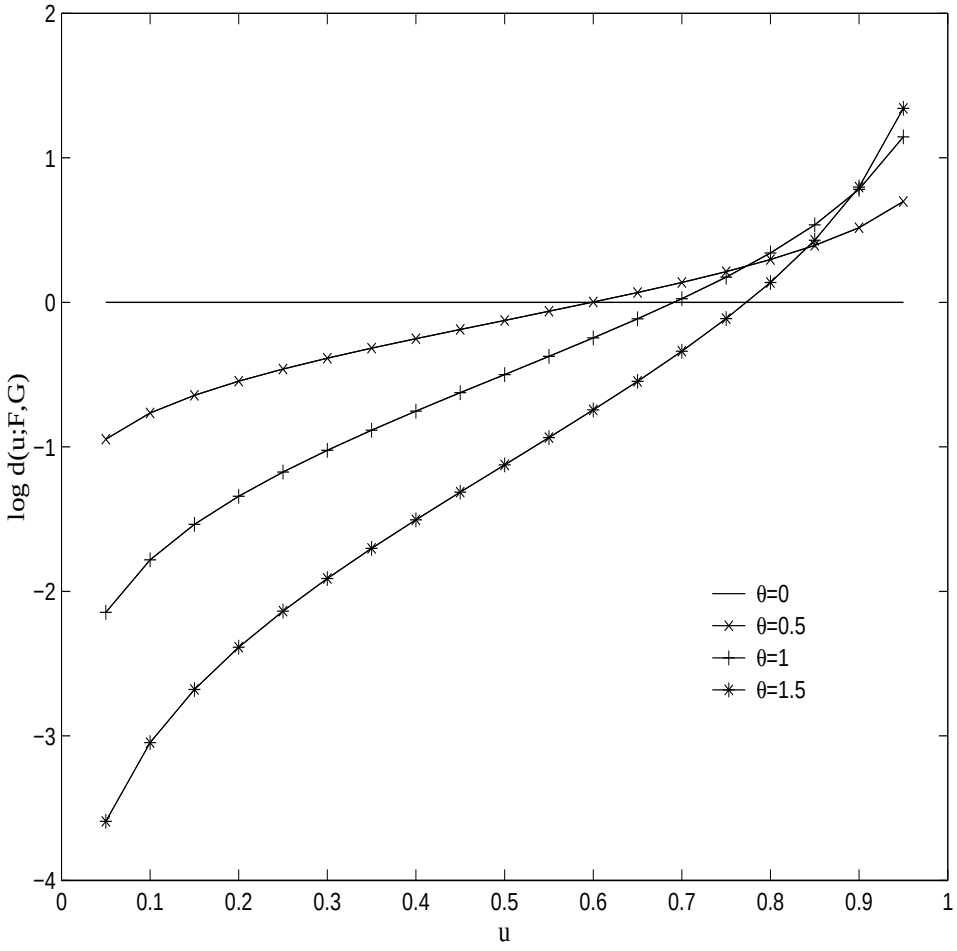

Figure 1. Plots of unpooled comparison density function of normal distributions with difference in locations assuming $F = N(0, 1)$ and $G = N(\theta, 1)$. Unpooled comparison density is $\log\, d(u; F, G) = \left(\theta\Phi^{-1}(u) - .5\theta^2\right)$ where $\Phi^{-1}(u)$ is the inverse function of $F = N(0, 1)$. As the difference in locations is getting greater, comparison density is getting farther from $\log\, d(u) = 0 (d(u) = 1)$. Also, $\log\, d(u)$ is monotone.

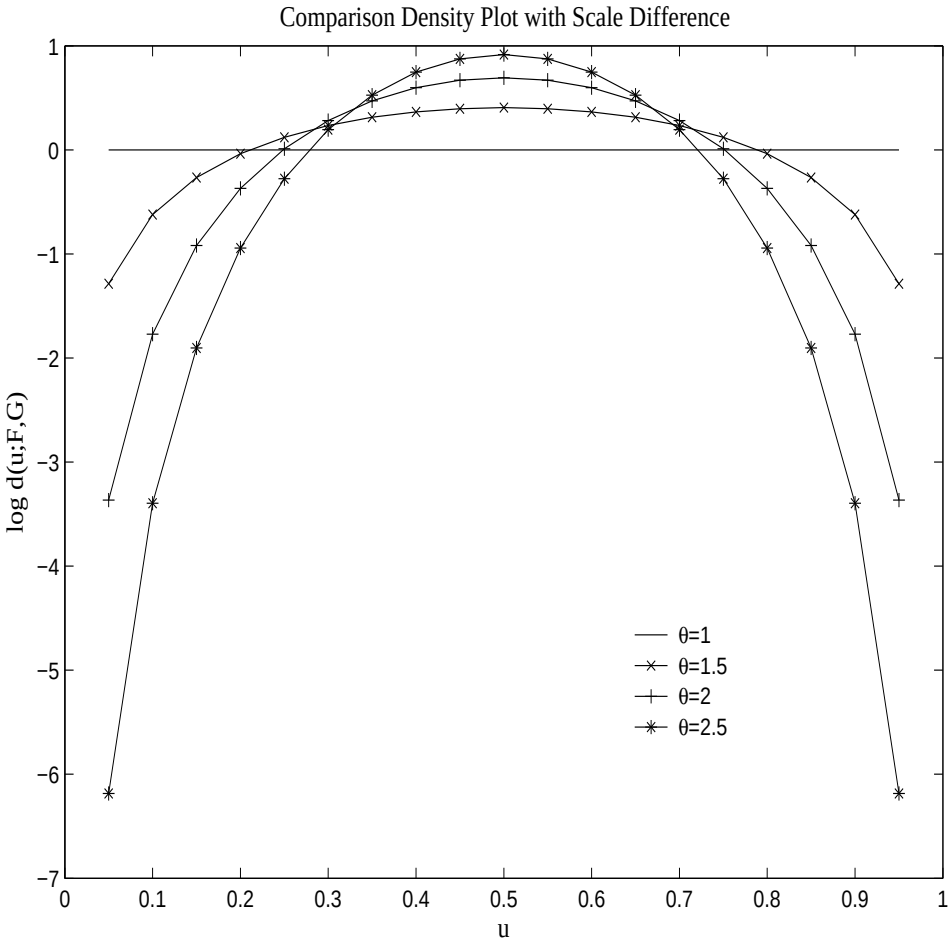

Figure 2. Plots of unpooled comparison density function of normal distributions with difference in scales assuming $F = N(0,1)$ and $G = N(0,\theta^{-2})$. Unpooled comparison density is $\log\, d(u;F,G) = \log\,\theta - .5\big(\Phi^{-1}(u)\big)^2(\theta^2-1)$ where $\Phi^{-1}(u)$ is the inverse function of $F = N(0,1)$. As the difference in scales is getting greater, comparison density is getting farther from $d(u) = 1(\log\, d(u) = 0)$. Also, $\log\, d(u)$ is quadratic and symmetric.

of unbounded functions is more difficult, we recommend to use pooled comparison density function $d(u)$ for estimation of a likelihood ratio.

Comparison distribution and comparison density concepts can be used to compare two discrete distributions F and G with respective probability mass functions p_F and p_G. We define a comparison density function

$$d(u) = d(u; F, G) = p_G(F^{-1}(u))/p_F(F^{-1}(u)); \tag{2.11}$$

then define a unpooled comparison distribution function

$$D(u) = D(u; F, G) = \int_0^u d(t)dt \tag{2.12}$$

assuming $p_F(x) = 0$ implies $p_G(x) = 0$. $D(u; F, G)$ is piecewise linear between its values at $u_j = F(x_j)$, where $x_1 < \cdots < x_m$ and $D(u_j; F, G) = G(F^{-1}(u_j)) = G(x_j)$.

Pooled comparison distribution and comparison density functions are defined as

$$d(u) \;=\; d(u; H, F) = p_F(H^{-1}(u))/p_H(H^{-1}(u)),$$

$$D(u) \;=\; D(u; H, F) = \int_0^u d(t)dt. \tag{2.13}$$

assuming $p_F(x) = 0$ implies $p_H(x) = 0$. $D(u; H, F)$ is piecewise linear between its values at $u_j = H(z_j)$, where $z_1 < \cdots < z_N$ and $D(u_j; H, F) = F(H^{-1}(u_j)) = F(z_j)$.

The graph of a comparison distribution $D(u; F, G)$ or $D(u; H, F)$ is called a P-P plot because it is a plot of $\big(F(y), G(y)\big)$ or $\big(H(y), F(y)\big)$ which compares the p values of an observation y under the two distributions. A P-P plot can be drawn by linearly connecting the points $(0, 0)$, $(1, 1)$, $\big(F(y), G(y)\big)$ or $\big(H(y), F(y)\big)$ for F-exact $u_j = F(y_j)(j = 1, \cdots, m)$ or H-exact $u_j = H(y_j)(j = 1, \cdots, N)$ respectively. And this is equal to the definition of $D(u)$ in discrete case. By using P-P plot, we

have a continuous sample distribution and can overcome the problem that sample distributions are discrete. P-P plot can be used as an analysis tool and provide the basis of further analysis.

2.3. Sample Comparison Distribution and Comparison Density Functions

For theoretical concepts to be applied to data analysis, it is crucial to define sample version of those concepts. Assume that we have a sample $X_1, \cdots, X_m$ from a continuous distribution F and a sample $Y_1, \cdots, Y_n$ from a continuous distribution G. Let $Z_1, \cdots, Z_N$ be a pooled sample of $X_1, \cdots, X_m$ and $Y_1, \cdots, Y_n$. Assume F and G have continuous densities f and g. Let $N = m + n$ and $\lambda_N = m/N$. Define sample distribution functions

$$\begin{aligned} F_m(x) &= \sum_{t=1}^{m} I(X_t \leq x)/m \\ G_n(y) &= \sum_{t=1}^{n} I(Y_t \leq y)/n \\ H_N(z) &= \lambda_N F_m(z) + (1-\lambda_N) G_n(z), \quad x, y, z \in R \end{aligned} \tag{2.14}$$

where $I(Y \leq y) = 1$ if $Y \leq y$ and $I(Y \leq y) = 0$ otherwise. H_N is a sample pooled distribution of F_m and G_n. The sample pooled distribution $H_N = \lambda_N F_m + (1-\lambda_N) G_n$ is equivalent to computing $H_N(y) = \sum_{t=1}^{m} I(X_t \leq y)/N + \sum_{t=1}^{n} I(Y_t \leq y)/N$.

Example: For illustration of concepts in this section on interesting data, we use a dataset from Giampaoli and Singer (2004) and call it as stress data. They consider the problem of comparing the mean of diastolic blood pressure of two group of individuals. One group is exposed to a stress stimulus (like the death of a close relative or discharge from employment) and another group is under normal conditions. The data are reproduced in Table 1 and each data value corresponds to the average of series of 30 measurements taken over periods of one hour to eliminate short term

Table 1

Diastolic blood pressure (mmHg)

Stress group	Control group
87.1	81.5
89.6	81.7
92.2	85.5
92.2	88.9
92.2	89.4
92.4	89.9
92.7	93.5
95.0	94.6
96.4	95.4
96.8	95.5
109.2	97.0

fluctuations. In this data, there are ties which means one value occurs several times like 92.2 in stress group. We consider F as distribution function of control group and G as distribution function of stress group. With this stress data, we compute sample distribution functions(Table 2- Table 4) using equaiton (2.14). Now we have $X_1, \cdots, X_{11}(m = 11)$ for control group, and $Y_1, \cdots, Y_{11}(n = 11)$ for stress group and thus $N = m + n = 22$. Thus $\lambda = m/N = 11/22 = 1/2$.

The sample unpooled comparison distribution is defined as a continuous function of u by

$$D^{\sim}(u; F, G) \quad = \quad G_n(F_m^{-1}(u)) \tag{2.15}$$

at u equal to F-exact value $u_j(j = 1, \cdots, m)$ satisfying $F_m(x_j) = u_j$ for distinct x_j values(Table 2). At other values of u, define $D^{\sim}(u; F, G)$ by linear interpolation between its values at F-exact values of u_j. Figure 3 and 4 are the plots of sample unpooled comparison distribution. Sample unpooled comparison distribution functions are defined by $D^{\sim}(u; F = Control, G = Stress)$ and $D^{\sim}(u; G = Stress, F = Control)$

Table 2

Sample distribution function for control group : $F_m(x) = \sum_{t=1}^{m} I(X_t \leq x)/m$, $m = 11$

Blood pressure	F_m
81.5	0.0909
81.7	0.1818
85.5	0.2727
88.9	0.3636
89.4	0.4545
89.9	0.5455
93.5	0.6364
94.6	0.7273
95.4	0.8182
95.5	0.9091
97.0	1.0000

Table 3

Sample distribution function for stress group : $G_n(y) = \sum_{t=1}^{n} I(Y_t \leq y)/n$, $n = 11$. *The number in* () *means the number of occurrences of the corresponding observation.*

Blood pressure	G_n
87.1	0.0909
89.6	0.1818
92.2(3)	0.4545
92.4	0.5455
92.7	0.6364
95.0	0.7273
96.4	0.8182
96.8	0.9091
109.2	1.0000

Table 4

Sample pooled distribution function $H_N(z) = \sum_{t=1}^{m} I(X_t \leq z)/N$ $+ \sum_{t=1}^{n} I(Y_t \leq z)/N$, $N = 22$

Blood pressure	H_N
81.5	0.0455
81.7	0.0909
85.5	0.1364
87.1	0.1818
88.9	0.2273
89.4	0.2727
89.6	0.3182
89.9	0.3636
92.2(3)	0.5000
92.4	0.5455
92.7	0.5909
93.5	0.6364
94.6	0.6818
95.0	0.7273
95.4	0.7727
95.5	0.8182
96.4	0.8636
96.8	0.9091
97.0	0.9545
109.2	1.0000

respectively for each figure. Specially from Figure 3, we can know that property of comparison distribution($D(1) = 1$) mentioned in section 2.2 is not satisfied. Sample pooled comparison distribution is

$$D^{\sim}(u; H, F) \quad = \quad F_m(H_N^{-1}(u)) \tag{2.16}$$

at H-exact values $u_j (j = 1, \cdots, r)$ satisfying $H_N(z_j) = u_j$ for distinct z_j values(Table 4) and at other values of u by linear interpolation between its values at H-exact values of u. Figure 5 is a plot of the sample pooled comparison distribution function with stress data. From sample comparison distribution, we compute the

sample comparison density $d^{\sim}(u)$ which is used as an estimate of $d(u)$. Actually, the slope of the sample comparison distribution is the sample comparison density. For the unpooled case,

$$d^{\sim}(u; F, G) = \frac{D^{\sim}(u_j; F, G) - D^{\sim}(u_{j-1}; F, G)}{u_j - u_{j-1}} \quad if \quad u_{j-1} < u < u_j \tag{2.17}$$

where $u_0 = 0$, and $u_j = F_m(x_j)$,$(j = 1, \cdots, m)$. For the pooled case,

$$d^{\sim}(u; H, F) = \frac{D^{\sim}(u_j; H, F) - D^{\sim}(u_{j-1}; H, F)}{u_j - u_{j-1}} \quad if \quad u_{j-1} < u < u_j \tag{2.18}$$

where $u_0 = 0$, and $u_j = H_N(z_j)$ for $j = 1, \cdots, N)$. Since our main concern is the pooled comparison density function, we have only a plot of the sample pooled comparison density function. See Figure 6. A pattern in a sample comparison density function indicates direction of shape of the score function whose statistic will be significant and therefore we could conclude a proper model for the difference of the two distributions. From Figure 6, we can see a quadratic pattern or somewhat cubic pattern and this may indicate the difference in the direction of 2nd(scale diffeence) or 3rd order score function. From the side by side boxplot of Figure 7, we see some differences in scale between two groups. In the Chapter III, we will have a more precise conclusion using the exponential model approach to the two-sample problem.

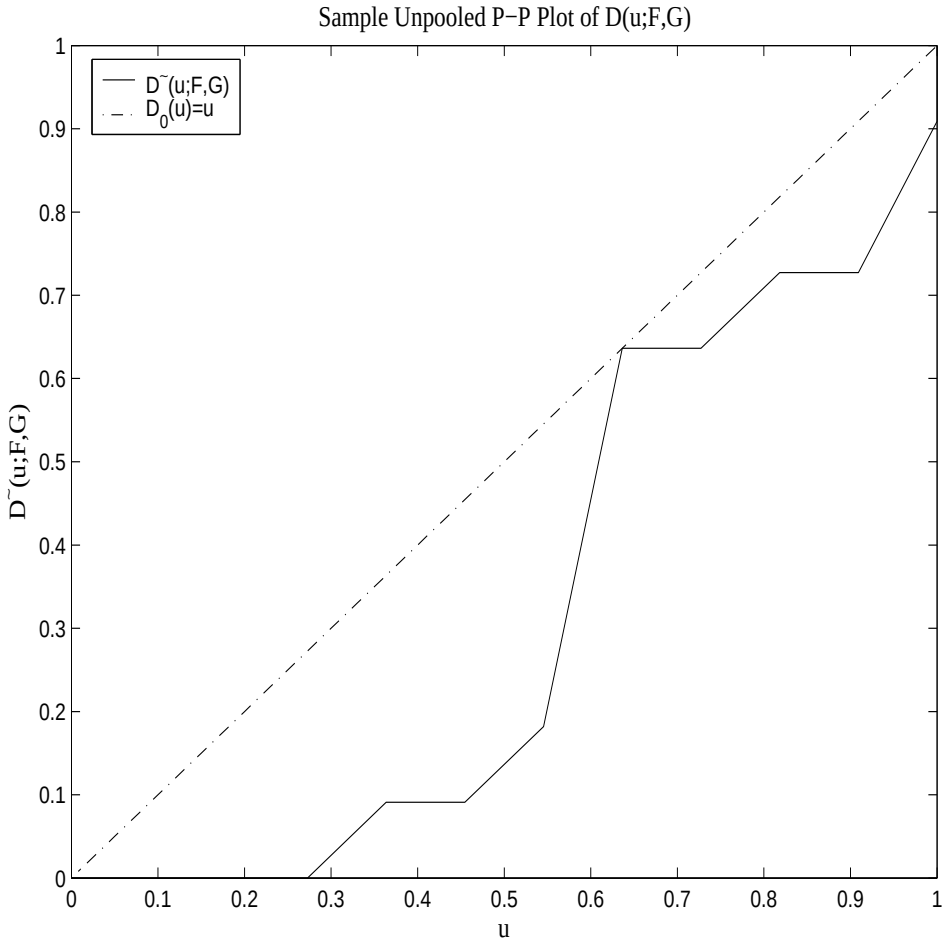

Figure 3. Plot of sample unpooled comparison distribution function $D^{\sim}(u; F, G)$ with stress data. In this case, two properties of comparison distribution function($D^{\sim}(0; G, F) = 0$ and ,$D^{\sim}(1; G, F) = 1$) are not satisfied.

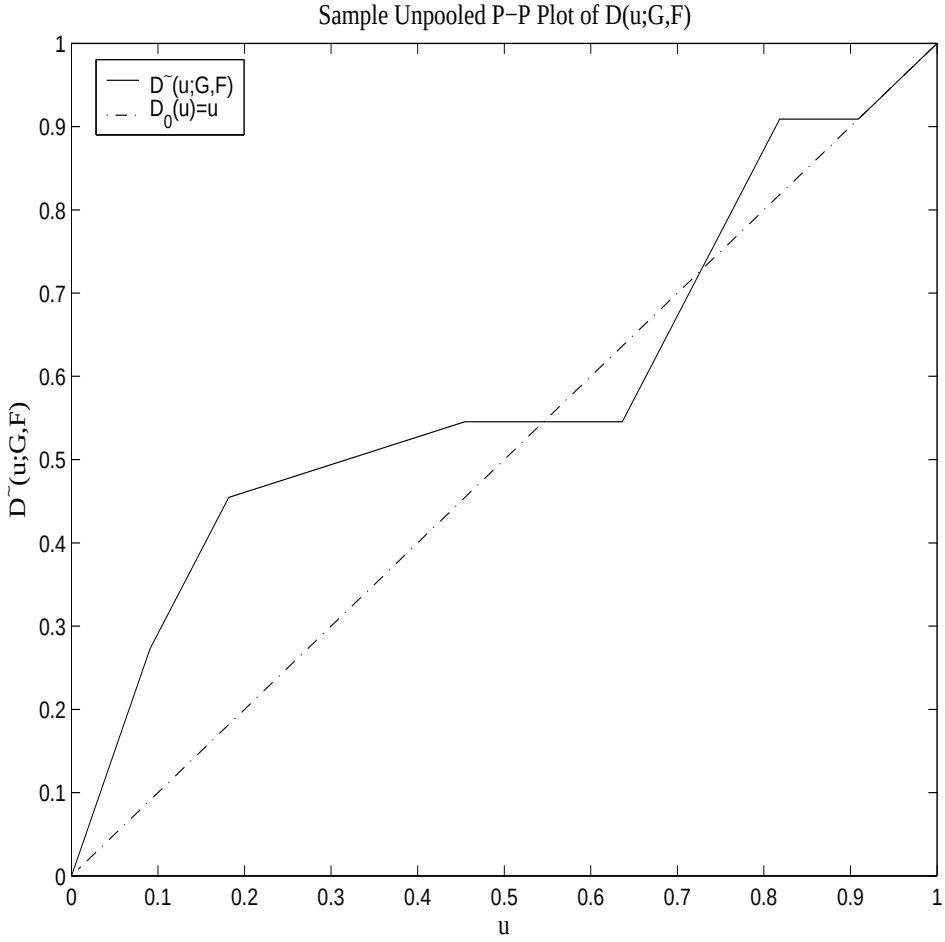

Figure 4. Plot of sample unpooled comparison distribution function $D^{\sim}(u; G, F)$ with stress data.

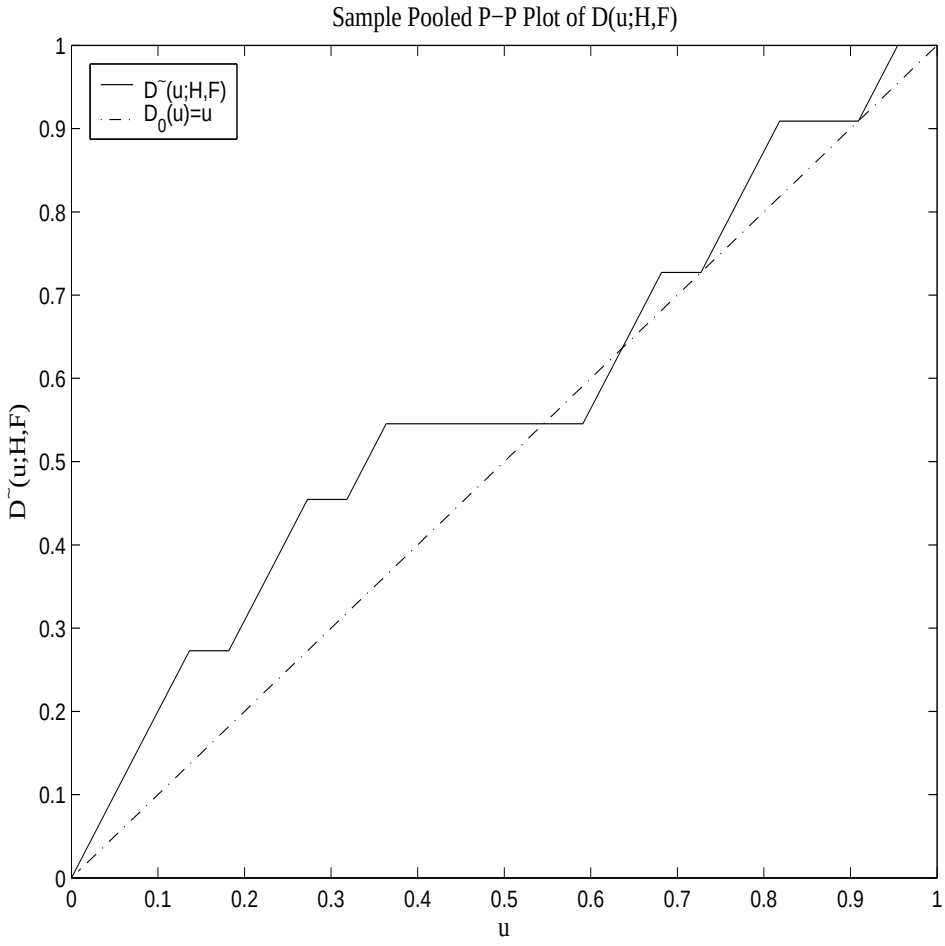

Figure 5. Plot of sample pooled comparison distribution function $D^{\sim}(u; H, F)$ with stress data.

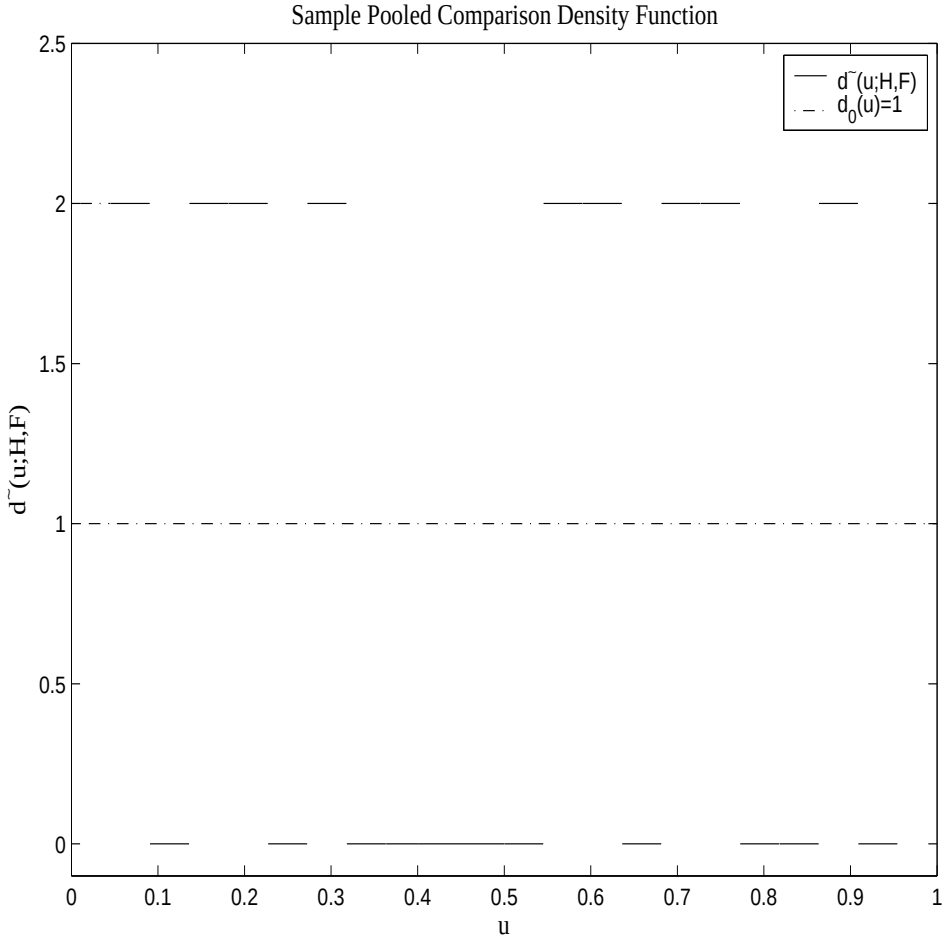

Figure 6. Plot of sample pooled comparison density function $d^{\sim}(u; H, F)$ with stress data.

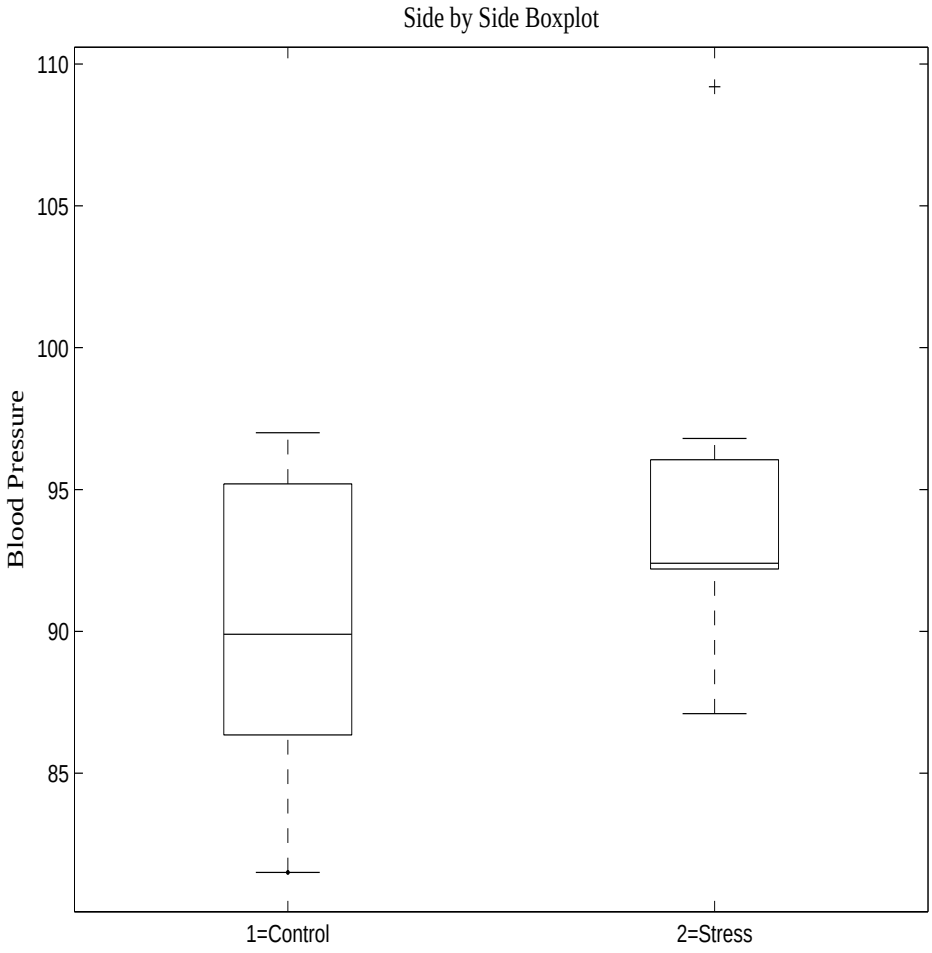

Figure 7. Side by side boxplot with stress data.

CHAPTER III

EXPONENTIAL MODEL WITH COMPONENTS

3.1. Introduction

In this chapter, we introduce an exponential model approach to the comparison density estimation and related concepts. To form an exponential model, we need to design mid-distribution score functions first. In subsection 3.2.1, we define the concept of mid-distribution function introduced by Parzen (1989) and in subsection 3.2.2, we provide a definition and recursive formula of mid-distribution score functions with an example.

To estimate the comparison density function, the exponential model with components will be used. For the estimation of comparison density $d(u)$, there have been two main approaches. One is kernel density estimation and another is autoregressive method. For details of each method, see Woodfield (1982) and Carmichael (1976) respectively. Our exponential model approach is similar to exponential family based density estimation, orthogonal series density estimation and maximum entropy method. Exponential family based density estimation is approximating a density function by using a member of a family of densities. Consider an exponential family of densities of the form

$$d(u;\boldsymbol{\theta}) = \exp\big(\sum_{k=1}^{K} \theta_k \phi_k(u) - \Psi_K(\boldsymbol{\theta})\big), \quad 0 < u < 1 \tag{3.1}$$

where $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_K) \in \Theta = \{\boldsymbol{\theta} \in \mathcal{R}^{\mathcal{K}} : \ominus_{\mathcal{K}}(\boldsymbol{\theta}) < \infty\}$. The function $\Psi_K(\boldsymbol{\theta})$ is the

normalizing value so that each density integrates to one:

$$\Psi_K(\boldsymbol{\theta}) = \log \left\{ \int_0^1 \exp\big(\sum_{k=1}^{K} \theta_k \phi_k(u)\big) du \right\} \tag{3.2}$$

and $\{\phi_k(u)\}_{k=1}^K$ are basis functions, which are bounded and linear independent functions such that

$$S_K = span\{1, \{\phi_k(u)\}_{k=1}^K\}$$

is a linear space. Three common choices for the basis functions are polynomials, trigonometric series, and spline bases. However disadvantage of this approach is the assumption that the comparison density is actually a member of this family which we do not assume in our exponential model approach.

Orthogonal series density estimation was introduced by Cencov (1962) and is allied with exponential family estimation. Cencov (1962) considers the expansion using orthogonal system with respect to a weight function. Basically Cencov's approach is a method of moments estimating scheme(Woodfield (1982)). Other researchers examined Cencov's method using specific system of orthogonal functions. Schwartz (1967) considers expansions using Hermite polynomials, Tarter and Kronmal (1970)) consider trigonometric systems(Fourier series expansion), and Crain (1974) uses Legendre polynomials. Consider the orthogonal series family of functions:

$$d(u; \boldsymbol{\theta}) = \theta_0 + \sum_{k=1}^{\infty} \theta_k \phi_k(u), \quad 0 < u < 1 \tag{3.3}$$

where $\theta_k \in \mathcal{R}$ and $\{\phi_k(u)\}_{k=1}^{\infty}$ form a complete orthonormal basis for the space of all square integrable functions on $[0, 1]$. Orthonormal means that

$$\int_0^1 \phi_i(u)\phi_j(u)du = I(i = j)$$

where $I(i = j) = 1$ if $i = j$ and 0 if $i \neq j$. By completeness of basis, there exists a

sequence of constants $\{\theta_k\}_{k=1}^{K}$ such that

$$|d(u) - \sum_{k=1}^{K} \theta_k \phi_k(u)|^2 \to 0$$

as $K \to \infty$. Thus, one can write

$$d(u) = \theta_0 + \sum_{k=1}^{\infty} \theta_k \phi_k(u), \quad 0 < u < 1$$

so that

$$\theta_k = \int_0^1 \phi_k(u) d(u) du = E\big(\phi_k(u)\big).$$

In practice, the comparison density might be estimated by

$$d^{\wedge}(u; \boldsymbol{\theta}) = \theta_0^{\wedge} + \sum_{k=1}^{K} \theta_k^{\wedge} \phi_k(u), \quad 0 < u < 1.$$

for suitable choice of order K and $\theta_k^{\wedge} = \sum_{j=1}^{m} \phi_k(R_j)/m$ where R_j is the rank of the sample in the pooled sample. Note that the estimator has the undesirable property that it may be negative for some value of u while our exponential model approach guarantees the nonnegativity of the estimate.

3.2. Basic Concepts

3.2.1. Mid-distribution Functions

Ranks of the observations are one of the important elements of nonparametric statistics. Parzen (1989) presented a concept of the mid-distribution function which is a transform of ranks. To compute mid-distribution score functions, we define mid-distribution functions first. Let F be a discrete distribution function. For distinct x values,

$$F^{mid}(x) = F(x) - .5 p_F(x) \tag{3.4}$$

where $F(x) = P[X \leq x]$ and $p_F(x) = P[X = x]$. For order statistics $X(1;m) \leq \cdots \leq X(m;m)$ of a sample $X_1, \cdots, X_m$ with no ties,

$$F^{mid}(X(j;m)) = \frac{(j-.5)}{m} = \frac{R_j - .5}{m} \tag{3.5}$$

which transforms the rank R_j to a number in the open unit interval, and is called mid-rank transform. If any X values are tied, their average rank is used for R_j. If X is a continuous random variable, $F^{mid}(X) \sim Uniform(0,1)$. This mid-distribution concept is important for discrete distributions, specially for sample distribution functions since sample distribution functions are discrete even though true distribution functions are continuous. For the mid-distribution transform $W = F^{mid}(X)$,

$$\begin{aligned} \mu_{mid} &= E\big(W\big) = 0.5 \\ \sigma^2_{mid} &= VAR\big(W\big) = [1 - E\big(p^2_F(X)\big)]/12. \end{aligned} \tag{3.6}$$

For equations in (3.6), there have been a few proofs and Parzen (2004) provides outline of a simple proof. For the proof of equations in (3.6), we adopt Parzen's approach. For detail, see appendix A.

In practice, assume that we have a sample $X_1, \cdots, X_m$. Then we estimate $F^{mid}(x)$ from

$$F^{mid}_m(x) = F_m(x) - .5p^{\sim}_F(x) \tag{3.7}$$

where $F_m(x) = \sum_{i=1}^m I(X \leq x)/m$ and $p^{\sim}_F(x) = 1/m$ with no ties. If there are ties, $p^{\sim}_F(x) = \sum_{t=1}^m I(X_t = x)/m$. Specially in the two-sample frame, let $Z_1, \cdots, Z_N$ be a pooled sample with a sample distribution function $H_N(z)$. Then, sample mid-distribution can be computed by

$$H^{mid}_N(z) = H_N(z) - .5p^{\sim}_H(z). \tag{3.8}$$

where $H_N^{mid}(z) = \sum_{t=1}^{N} I(Z_t \leq z)/N$ and $p_H^{\sim}(z) = \sum_{t=1}^{N} I(Z_t = z)/N$. With stress data, we compute $H_N(z)$, $p_H^{\sim}(z)$, and $H_N^{mid}(z)$ in Table 5.

3.2.2. Design of Score Functions

Let X be a variable with distribution function F. Then a score function ψ is defined satisfying

$$\begin{aligned} E\big(\psi(X)\big) &= 0, \\ VAR\big(\psi(X)\big) &= 1 \end{aligned} \tag{3.9}$$

where expectation is taken with respect to a specific distribution of X. For discrete F, we define orthonormal score functions which are based on ranks through mid-distribution transform. By Gram-Schmidt orthonormalization, we derive orthogonal polynomials, called mid-distribution score functions, recursively. Define $w_1(X)$, $\phi_1(X)$ and $\psi_1(X)$.

$$\begin{aligned} w_1(X) &= F^{mid}(X) - \mu_{mid}, \\ \phi_1(X) &= w_1(X), \\ \psi_1(X) &= \frac{\phi_1(X)}{\sqrt{\langle \phi_1(X), \phi_1(X) \rangle}} = \frac{F^{mid}(X) - \mu_{mid}}{\sigma_{mid}} \end{aligned} \tag{3.10}$$

where $\langle \cdot, \cdot \rangle$ is inner product of two functions, $\mu_{mid} = E\big(F^{mid}(X)\big) = 0.5$ and $\sigma_{mid}^2 = VAR\big(F^{mid}(X)\big)$ which are defined in the previous subsection 3.2.1. For $j = 2, 3, \cdots$, we have a recursive form

$$\begin{aligned} w_j(X) &= \psi_1^j(X) - \beta_j, \\ \phi_j(X) &= w_j(X) - \sum_{i=1}^{j-1} \langle \phi_j(X), \psi_i(X) \rangle \psi_i(X), \\ \psi_j(X) &= \frac{\phi_j(X)}{\sqrt{\langle \phi_j(X), \phi_j(X) \rangle}} \end{aligned} \tag{3.11}$$

where $\beta_r = E[\big((F^{mid}(X) - 0.5)/\sigma_{mid}\big)^r]$. A few terms of mid-distribution score functions $\psi_j(X)$ are derived as follows;

$$\begin{aligned} \psi_0(X) &= 1 \\ \psi_1(X) &= (F^{mid}(X) - .5)/\sigma_{mid} \\ \psi_2(X) &= \left[(\psi_1^2(X) - 1) - \beta_3\psi_1(X)\right]/a_2 \\ &\vdots \end{aligned} \tag{3.12}$$

where $a_2^2 = \beta_4 - \beta_3^2 - 1$. Also, the equations in (3.9) are satisfied by the orthonormality of ψ functions. In the two-sample work in practice, let $Z_1, Z_2, \cdots, Z_N$ be a pooled sample with a sample distribution $H_N(z)$. Then, the sample mid-distribution is computed by equation (3.8). With the stress data, we compute mid-distribution score functions up to order 4 by using the above recursive formula in equation(3.11). Figure 8 shows a plot of each sample mid-distribution score function. The plots are on the unit interval and plotting $\psi_j\big(H_N^{-1}(u)\big)$ for $u_{i-1} < u < u_i$ and $H_N^{-1}(u_i) = z_i$. Practically we can verify the orthornormality of sample mid-distribution score functions defined on the unit interval using stress data. See the Table 6.

3.2.3. Components

The usefulness of $d(u; H, F)$ comparing F and H arises from the fact that $d(u; H, F) = 1$ *iff* $H(y) = F(y)$. Thus one method to compare F and H can be based on a comparison of $d(u; H, F)$ with the uniform density $p_0(u) = 1$ when $0 \leq u \leq 1$. Eubank et al. (1987) gives the introduction of a measure of the disparity between $d(u; H, F)$ and $p_0(u)$ and analysis of its component decomposition. Define the measure using the squared $L_2[0, 1]$ norm of their differences

$$\phi^2 = \| \, d - p_0 \, \|^2 = \int_0^1 d(u)^2 du - 1 \tag{3.13}$$

Let $\{\psi_i\big(H^{-1}(u)\big)\}_{i=1}^{\infty}$ be an orthonormal system for $L_2[0,1]$ such that

$$d - p_0 \sim \sum_{j=1}^{\infty} \theta_j \psi_j\big(H^{-1}(u)\big) \tag{3.14}$$

where the θ_j's are generalized Fourier coefficients.

$$\theta_j = \int_0^1 (d(u)-1)\psi_j\big(H^{-1}(u)\big)du = \int_0^1 d(u)\psi_j\big(H^{-1}(u)\big)du, \quad j = 1, 2, \cdots \tag{3.15}$$

and $\sim$ is the usual Fourier series notation indicating $\sum_{j=1}^{r} \theta_j\psi_j \to d - p_0$ in $L_2[0,1]$ as $r \to \infty$. By Parseval's relation,

$$\phi^2 = \sum_{j=1}^{\infty} \theta_j^2 \tag{3.16}$$

where θ_j's are components of ϕ^2. Thus, to test $H_0 : F = H$ is equivalent to test $H_0 : \phi^2 = 0$ or to test $\theta_j = 0$ for all $j \geq 1$ from equation (3.16). Since one cannot test an infinite number of parameters, Eubank et al. (1987) suggest to test subhypotheses, such as $H_0 : \theta_j = 0$ for $j = 1, \cdots, M$ for a constant M. $H_0 : F = H$ should be rejected if we can reject any of the subhypotheses $\theta_j = 0 (j = 1, \cdots, M)$. To estimate θ_j, make the change of variable $x = H^{-1}(u)$ in equation(3.15). Then,

$$\begin{aligned}
\theta_j &= \int_0^1 d(u)\psi_j\big(H^{-1}(u)\big)du \\
&= \int_0^1 \frac{f\big(H^{-1}(u)\big)}{h\big(H^{-1}(u)\big)}\psi_j\big(H^{-1}(u)\big)du \\
&= \int_{-\infty}^{\infty} \frac{f(x)}{h(x)}\psi_j(x)h(x)dx \\
&= \int_{-\infty}^{\infty} \psi_j(x)f(x)dx \\
&= E_f(\psi_j(x))
\end{aligned} \tag{3.17}$$

Thus, given estimates F_m and H_N for F and H, the estimate of θ_j is

$$\begin{aligned} \theta_j^{\sim} &= \int_{-\infty}^{\infty} \psi_j(x) dF_m(x) \\ &= \sum_{i=1}^{m^*} p_F^{\sim}(x_i^*)\psi_j(x_i^*) \\ &= E_F(\psi_j(x)) \end{aligned} \tag{3.18}$$

where x_i^* is the distinct value of the first sample X_i in the pooled sample of X and Y and m^* is the number of distinct values in the sample of X and $p_F^{\sim}(x_i^*) = \sum_{i=1}^{m^*} I(X = x_i^*)/m$. To test $H_0 : \theta_j = 0$, we need to know asymptotic distribution of the individual components. For that, the corollary of Eubank et al. (1987) is used, which is a variant of the Chernoff-Savage theorem(Chernoff and Savage (1958)). Chernoff and Savage (1958) define a rank statistic having the form

$$\begin{aligned} S_N &= \int_{-\infty}^{\infty} J_N\big(H_N(x)\big) dF_m \\ &= \frac{1}{m} \sum_{j=1}^{m} J_N(R_i/N) \end{aligned} \tag{3.19}$$

where R_i is the rank of X_i in the pooled sample, J_N is known as a score function, and F_m and H_N are sample distributions defined in Chapter II. Using the following Chernoff-Savage approach, we demonstrate normality of $\theta_j^{\sim}$.

THEOREM 3.1. *(Chernoff and Savage, 1958). If $J(u)$ is not constant and if $|J^{(i)}| \leq K|u(1-u)|^{-i-(1/2)+\delta}$ for $i = 0, 1, 2$ and some K and $\delta > 0$, then for fixed and continuous F and G, one has S_N is $AN(\mu, \sigma_N^2)$, where*

$$\mu = \int J\big(H(x)\big) dF(x) \tag{3.20}$$

and

$$\begin{aligned} N\sigma_N^2 &= 2(1-\lambda_N)\{\iint_{x<y} G(x)\big(1-G(y)\big)J'\big(H(x)\big)J'\big(H(y)\big)dF(x)dF(y) \\ &+ \frac{1-\lambda_N}{\lambda_N}\iint_{x<y} F(x)\big(1-F(y)\big)J'\big(H(x)\big)J'\big(H(y)\big)dG(x)dG(y)\} \quad (3.21) \end{aligned}$$

providing $\sigma_N \neq 0$.

The notation S_N is $AN(\mu, \sigma_N^2)$ means that the distribution function of the random variable $(S_N - mu)/\sigma_N$ converges pointwise to the distribution function of a standard normal random variable. To find approximate values of the distribution function of S_N, one need only calculate the values of μ and σ_N^2. In many practical circumstances, the values of μ and σ_N^2 can be worked out. For an example, see Alexander (1989).

$\theta_j^{\sim}$'s asymptotic distribution under the null hypothesis $F = G$ or $\theta_j = 0$, $j = 1, 2, \cdots$ is given in the following theorem.

THEOREM 3.2. *Under* $H_0 : \theta_j = 0$, *the asymptotic distribution of* $\sqrt{N}\theta_j^{\sim}$ *is* $N(0, \sigma_j^2)$ *where*

$$\sigma_j^2 = \frac{1-\lambda}{\lambda}\int_0^1 \psi_j^2(H^{-1}(u))du = \frac{1-\lambda}{\lambda}.$$

For the proof, see Alexander (1989). And σ_j^2 is estimated by

$$\sigma_j^{\sim 2} = \frac{1-\lambda_N}{\lambda_N}.$$

Then we find significant components by testing $H_0 : \theta_j = 0$, $j = 1, \cdots, M$ through standardization. Since we know the asymptotic distribution of $\theta_j^{\sim}$ which is an estimate of θ_j from Theorem 3.2, if the result of standardization is greater than 2(or 3) in absolute value, we conclude that θ_j is not zero or significant(or very significant). With stress data, we have $\theta_j^{\sim}$ values up to order 4 and corresponding standardized

values defined as $C_j = \sqrt{N}(\theta_j^{\sim})/\sigma_j^{\sim}$

$$\theta_1^{\sim} = -0.2367 \qquad \sigma_1^{\sim} = 1$$
$$\theta_2^{\sim} = 0.2456 \qquad \sigma_2^{\sim} = 1$$
$$\theta_3^{\sim} = -0.2591 \qquad \sigma_3^{\sim} = 1$$
$$\theta_4^{\sim} = -0.2088 \qquad \sigma_4^{\sim} = 1$$

then,

$$\begin{aligned} C1 &= -1.1102, \\ C2 &= 1.1520, \\ C3 &= -1.1253, \\ C4 &= -0.9794. \end{aligned}$$

Thus, one may be able to conclude that there are no significant components through the test results. Therefore there is not enough evidence to reject $H_0 : \theta_j = 0$ for $j = 1, 2, 3, 4$. This could mean that $d(u; H, F)$ is not different from $p_0(u) = 1$ and we could conclude that there is no significant evidence to reject $H_0 : F = G$.

3.2.4. Exponential Model Approach and Comparison Density Estimation

The model discussed in this dissertation is motivated by the observation that the logarithm of a probability function is often found to be a fairly well-behaved function and it is often convenient to work with it. In terms of density estimation, the exponential model guarantees the nonnegativity of the density function which is an essential property of a density function.

Our exponential model is formed using score functions which have largest components instead of finding significant components through tests performed in the previ-

ous subsection 3.2.3. With selected components and corresponding mid-distribution score functions, we form an exponential model estimator of comparison density function:

$$d^{\wedge}(u;\boldsymbol{\theta}) = \exp\big(\theta_0^{\wedge} + \sum_{k\in\mathcal{C}} \theta_k^{\wedge}\psi_k\big(H^{-1}(u)\big)\big) \tag{3.22}$$

where $\mathcal{C}$ is a set of index of selected components. With stress data, we sort $\theta_j^{\sim}$ values in absolute value in descending order.

$$\begin{aligned}
\theta_3^{\sim} &= -0.2591 \\
\theta_2^{\sim} &= 0.2456 \\
\theta_1^{\sim} &= -0.2367 \\
\theta_4^{\sim} &= -0.2088
\end{aligned}$$

Then with the three two components, an exponential model estimator of is

$$d^{\wedge}(u;\boldsymbol{\theta}) = \exp\big(\theta_0^{\wedge} + \sum_{\mathcal{C}} \theta_k^{\wedge}\psi_k\big(H^{-1}(u)\big)\big) \tag{3.23}$$

where $\mathcal{C} = \{1, 2, 3\}$. Estimation of $\theta_k^{\wedge}$ will be discussed in 3.2.4.2.

3.2.4.1. Maximum Entropy Interpretation of Exponential Model Approach

Our comparison density estimator using exponential model approach has a maximum entropy interpretation in the sense that maximum entropy density estimation gives the same form of estimator as exponential model estimator. However, our exponential model is different in the sense that we use orthonormal score functions as basis functions and use significant terms.

Shannon's measure of entropy was originally developed for a discrete distribution. The entropy of a discrete distribution, denoted by $H_S(\cdot)$ is defined

$$H_S(p) = -\sum_x p(x) \log\, p(x) \tag{3.24}$$

where $p(x)$ is probability mass function. The notion of entropy for a continuous distribution is formally defined by

$$H_S(f) = -\int_{-\infty}^{\infty} f(y) \log\, f(y) dy \tag{3.25}$$

with probability density function $f(y)$. Another fundamental concept is cross-entropy defined by

$$H(f; g) = \int_{-\infty}^{\infty} \big(-\log\, g(y)\big) f(y) dy \tag{3.26}$$

A closely related concept is Kullback-Leibler's information divergence $I(f; g)$ between two probability density functions $f(y)$ and $g(y)$. The information divergence is defined by

$$I(f : g) = \int_{-\infty}^{\infty} \left(-\log\, \frac{g(y)}{f(y)}\right) f(y) dy. \tag{3.27}$$

Then one important property of $I(f; g)$ is

$$I(f; g) \geq 0 \tag{3.28}$$

with equality *if and only if* $f = g$, which is called Shannon's inequality.

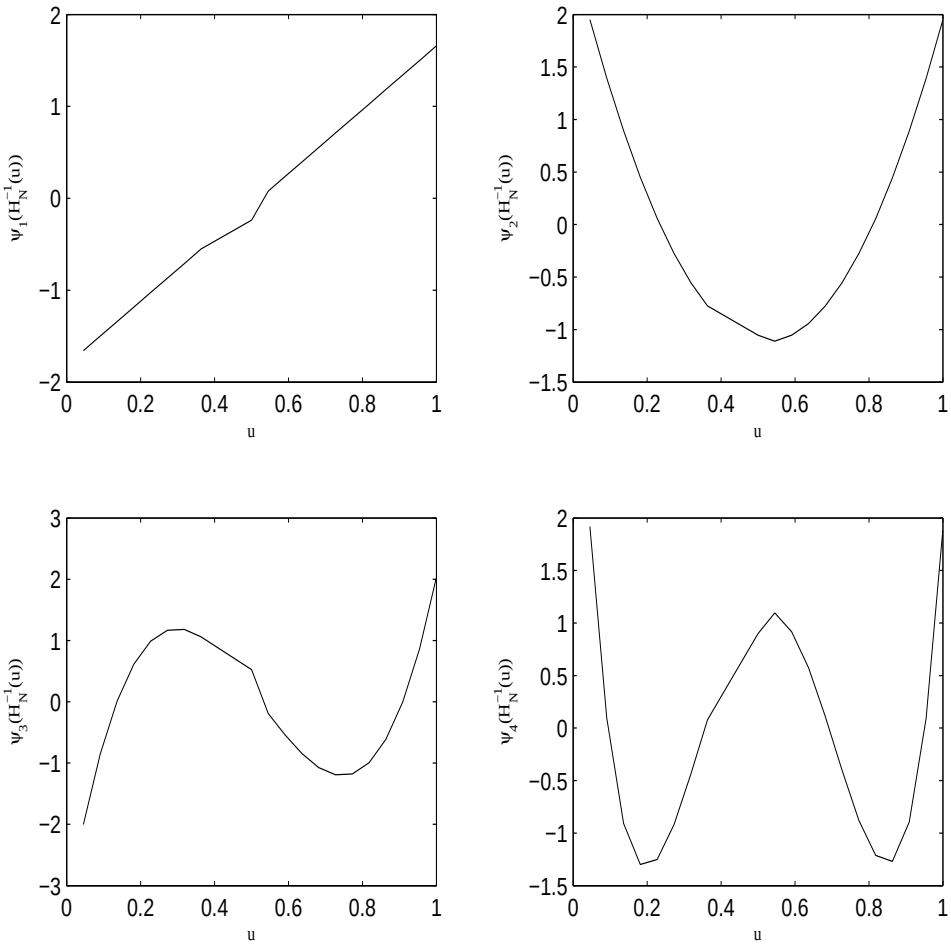

Figure 8. Sample mid-distribution score functions up to order 4 using stress data. $\psi_j\big(H_N^{-1}\big(u\big)\big)$ for $u_{i-1} < u < u_i$ and $H_N^{-1}(u_i) = z_i$.

Exponential model for comparison density $d(u; H, F)$ has a form

$$\log\, d(u; H, F) = \sum_{j=1}^{K} \theta_j \psi_j\big(H^{-1}(u)\big) - \Psi_0(\boldsymbol{\theta}) \tag{3.29}$$

where

$$\Psi_0(\boldsymbol{\theta}) = \log \int_0^1 e^{\sum_{j=1}^{K} \theta_j \psi_j\left(H^{-1}(u)\right)} du \tag{3.30}$$

and $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_K)$.

THEOREM 3.3. *Among any comparison density $d(u)$ satisfying the following constraints*

$$\int_0^1 \psi_j\big(H^{-1}(u)\big) d(u) du = \tau_j, \quad j = 1, \cdots, M,$$

an exponential model $d_0(u)$ has maximum entropy.

proof:

$$\begin{aligned} I\big(d(u); d_0(u)\big) &= \int_0^1 \left(-\log \frac{d_0(u)}{d(u)}\right) d(u) du \\ &= \int_0^1 \big(-\log\, d_0(u)\big) d(u) du + \int_0^1 \big(\log\, d(u)\big) d(u) du \\ &= H_S\big(d(u); d_0(u)\big) - H_S\big(d(u)\big) \end{aligned} \tag{3.31}$$

$$\begin{aligned} H(d(u); d_0(u)) &= \int_0^1 \big(-\log\, d_0(u)\big) d(u) du \\ &= \int \left[-\sum_{j=1}^{M} \theta_j \psi_j\big(H^{-1}(u)\big) - \Psi_0(\boldsymbol{\theta})\right] d(u) du \\ &= -\Psi_0(\boldsymbol{\theta}) - \sum_{j=1}^{M} \theta_j \tau_j \\ &= H_S\big(d_0(u)\big) \end{aligned} \tag{3.32}$$

Thus,

$$\begin{aligned} I\big(d_0(u); d(u)\big) &= H_S\big(d(u); d_0(u)\big) - H_S\big(d_0(u)\big) \geq 0 \quad \text{by} \quad \text{equation(3.28)} \\ &\Rightarrow \quad H_S\big(d_0(u)\big) \geq H_S\big(d(u)\big) \end{aligned} \tag{3.33}$$

Therefore, an exponential model of comparison density function $d(u)$ has maximum entropy.

3.2.4.2. Estimation of Coefficients of An Exponential Model

To estimate coefficients of $d^{\wedge}(u)$ of equation (3.22), we adopt the method of moments. The method of moments is a technique for constructing estimators that is based on matching the sample moments with the corresponding distribution moments. Let

$$\mu_k^{\sim}(\boldsymbol{\theta}) = \int_0^1 d^{\sim}(u)\psi_k\big(H^{-1}(u)\big)du \tag{3.34}$$

denote the k^{th} sample moment where $k = 1, 2, \cdots, K$. Next, let

$$\mu_k(\boldsymbol{\theta}) = \int_0^1 d^{\wedge}(u)\psi_k\big(H^{-1}(u)\big)du \tag{3.35}$$

denote the k^{th} moment. To construct estimators of coefficients of exponential model, we need to solve the set of simultaneous equations

$$\begin{aligned} \mu_1^{\sim}(\boldsymbol{\theta}) &= \mu_1(\boldsymbol{\theta}), \\ \mu_2^{\sim}(\boldsymbol{\theta}) &= \mu_2(\boldsymbol{\theta}), \\ &\vdots \\ \mu_K^{\sim}(\boldsymbol{\theta}) &= \mu_K(\boldsymbol{\theta}). \end{aligned} \tag{3.36}$$

Then equations in (3.36) can be rewritten by

$$M_k(\boldsymbol{\theta}) = \int_0^1 \big(d^{\wedge}(u) - d^{\sim}(u)\big)\psi_k\big(H^{-1}(u)\big)du = 0, \quad k = 1, 2, \cdots, K, \tag{3.37}$$

where $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_K)'$ and those satisfying constraints $M_k(\boldsymbol{\theta}) = 0$ have a maximum entropy interpretation from theorem 3.3. Assume we have 4 components θ_k, $k = 1, 2, 3, 4$. Then the solutions are updated according to the scheme from Newton-Raphson method

$$\boldsymbol{\theta}^{(i+1)} = \boldsymbol{\theta}^{(i)} + \Delta\boldsymbol{\theta}^i,$$

where i indicates the iteration number and $\Delta\boldsymbol{\theta} = (\Delta\theta_1, \Delta\theta_2, \Delta\theta_3, \Delta\theta_4)'$. We have the Jacobian system with starting values $\theta_k^{(0)}$, $k = 1, 2, 3, 4$.

$$\begin{pmatrix} \frac{\partial M_1(\boldsymbol{\theta})}{\partial\theta_1} & \frac{\partial M_1(\boldsymbol{\theta})}{\partial\theta_2} & \frac{\partial M_1(\boldsymbol{\theta})}{\partial\theta_3} & \frac{\partial M_1(\boldsymbol{\theta})}{\partial\theta_4} \\ \frac{\partial M_2(\boldsymbol{\theta})}{\partial\theta_1} & \frac{\partial M_2(\boldsymbol{\theta})}{\partial\theta_2} & \frac{\partial M_2(\boldsymbol{\theta})}{\partial\theta_3} & \frac{\partial M_2(\boldsymbol{\theta})}{\partial\theta_4} \\ \frac{\partial M_3(\boldsymbol{\theta})}{\partial\theta_1} & \frac{\partial M_3(\boldsymbol{\theta})}{\partial\theta_2} & \frac{\partial M_3(\boldsymbol{\theta})}{\partial\theta_3} & \frac{\partial M_3(\boldsymbol{\theta})}{\partial\theta_4} \\ \frac{\partial M_4(\boldsymbol{\theta})}{\partial\theta_1} & \frac{\partial M_4(\boldsymbol{\theta})}{\partial\theta_2} & \frac{\partial M_4(\boldsymbol{\theta})}{\partial\theta_3} & \frac{\partial M_4(\boldsymbol{\theta})}{\partial\theta_4} \end{pmatrix} \begin{pmatrix} \Delta\theta_1 \\ \Delta\theta_2 \\ \Delta\theta_3 \\ \Delta\theta_4 \end{pmatrix} = - \begin{pmatrix} M_1(\boldsymbol{\theta}) \\ M_2(\boldsymbol{\theta}) \\ M_3(\boldsymbol{\theta}) \\ M_4(\boldsymbol{\theta}) \end{pmatrix}$$

which is obtained by Taylor expansion with

$$\frac{\partial M_k(\boldsymbol{\theta})}{\partial\theta_l} = \int_0^1 \exp\left[\theta_0 + \sum_{j=1}^{4}\theta_j\psi_j\big(H^{-1}(u)\big)\right]\psi_k\big(H^{-1}(u)\big)\psi_l\big(H^{-1}(u)\big)du. \tag{3.38}$$

In practice, initial values are computed from

$$\theta_k^{(0)} = \sum_{i=1}^{m^*} p_F^{\sim}(x_i^*)\psi_j(x_i^*) \quad k = 1, 2, 3, 4 \tag{3.39}$$

which is from equation (3.18). Table 7 provides the result of Newton-Raphson iteration to estimate the components of exponential model using stress data. Three components was chosen from subsection 3.2.4. Then, the estimated exponential model

is

$$d^{\wedge}(u;\boldsymbol{\theta}) = \exp\big(-0.052-0.2171\psi_1\big(H^{-1}(u)\big)+0.0922\psi_2\big(H^{-1}(u)\big)-0.1862\psi_3\big(H^{-1}(u)\big)\big). \tag{3.40}$$

From the estimated coefficients, we could know that two distributions may be different in the direction of 1st and 3rd score functions.

To check the goodness of fit of our exponential model, we compute a smooth comparison distribution function and this wil be discussed in Chapter IV.

Table 5

Sample mid-distribution function $H_N^{mid}(z)$ *:* $H_N^{mid}(z) = H_N(z) - .5p_H^{\sim}(z)$

Blood pressure	R_j	H_N	$p_H^{\sim}(z)$	$H_N^{mid}(z)$
81.5	1	0.0455	0.0455	0.0227
81.7	2	0.0909	0.0455	0.0682
85.5	3	0.1364	0.0455	0.1136
87.1	4	0.1818	0.0455	0.1591
88.9	5	0.2273	0.0455	0.2045
89.4	6	0.2727	0.0455	0.2500
89.6	7	0.3182	0.0455	0.2955
89.9	8	0.3636	0.0455	0.3409
92.2(3)	10	0.5000	0.1364	0.4318
92.4	12	0.5455	0.0455	0.5227
92.7	13	0.5909	0.0455	0.5682
93.5	14	0.6364	0.0455	0.6136
94.6	15	0.6818	0.0455	0.6591
95.0	16	0.7273	0.0455	0.7045
95.4	17	0.7727	0.0455	0.7500
95.5	18	0.8182	0.0455	0.7955
96.4	19	0.8636	0.0455	0.8409
96.8	20	0.9091	0.0455	0.8864
97.0	21	0.9545	0.0455	0.9318
109.2	22	1.0000	0.0455	0.9773

Table 6

Inner product of score functions to verify orthonormality with stress data

	$\psi_1(H_N^{-1}(u))$	$\psi_2(H_N^{-1}(u))$	$\psi_3(H_N^{-1}(u))$	$\psi_4(H_N^{-1}(u))$
$\psi_1(H_N^{-1}(u))$	1.000	-5.55-11e-017	4.9960e-016	-6.3838e-016
$\psi_2(H_N^{-1}(u))$	-5.55-11e-017	1.000	0	1.3045e-015
$\psi_3(H_N^{-1}(u))$	4.9960e-016	0	1.000	0
$\psi_4(H_N^{-1}(u))$	-6.3838e-016	1.3045e-015	0	1.000

Table 7

$\theta_j^{\wedge}$ values up to order 3 through Newton-Raphson iteration with stress data

Iteration	θ_0	θ_1	θ_2	θ_3
1	-0.1210	-0.2367	0.2456	-0.2591
2	-0.0667	-0.2354	0.1136	-0.2102
3	-0.0539	-0.2198	0.0949	-0.1896
4	-0.0522	-0.2174	0.0925	-0.1865
5	-0.0520	-0.2172	0.0923	-0.1862
6	-0.0520	-0.2171	0.0922	-0.1862
7	-0.0520	-0.2171	0.0922	-0.1862

CHAPTER IV

TWO-SAMPLE DATA ANALYSIS PROCEDURE

4.1. Introduction

In this chapter, we implement a two-sample data analysis procedure based on exponential model approach introduced at Chapter III. Section 4.2 gives our two-sample data analysis algorithm based on exponential model approach and to illustrate each step, the stress dataset is used again.

4.2. Algorithm

Step 1: Combine two samples and arrange them in order. Estimate comparison distribution $D(u; H, F)$ as $D^{\sim}(u_j; H, F)$

$$\begin{aligned} D^{\sim}(u_j; H, F) &= F_m\big(H_N^{-1}(u_j)\big) = F_m(z_j) \\ &= \sum_{t=1}^{m} I(X_t \le H_N^{-1}(u_j))/m. \end{aligned} \tag{4.1}$$

As an estimate of $D(u; H, F)$, we use a P-P plot drawn by connecting points $(0, 0)$, $(u_j, D^{\sim}(u_j; H, F))$, and $(1, 1)$ where $u_j = H_N(z_j)$ called H-exact values for distinct z_j values in pooled sample and $D^{\sim}(u_j; H, F) = F_m(H_N^{-1}(u_j))$. If P-P plot is close to the 45 degree straight line, that implies $F = H$. However, since $D^{\sim}(u; H, F)$ is usually very rough, it is not easy to make a conclusion. Thus, we need to estimate a smooth comparison distribution $D^{\wedge}(u; H, F)$. To estimate this, go to step 2. Figure 5 is a

P-P plot using stress data.

Step 2: Compute mid-distribution score functions $\psi_j\big(H_N^{-1}(u)\big)$.

At distinct values in the pooled sample, compute mid-distribution score functions using Gram-Schmidt orthonormalization up to order 4. Use the recursive formula in equations (3.10) and (3.11). Table 8 shows mid-distribution score functions using stress data.

Step 3: Compute the estimated values of components $\theta_j^{\sim}$. Select largest $\theta_j^{\sim}$.

$$\theta_j^{\sim} = \sum_{i=1}^{m^*} p_F^{\sim}(x_i^*)\psi_j(x_i^*). \tag{4.2}$$

which is defined at equation (3.18). Here m^* is the number of distinct values x_i^* from the first sample. Select largest $\theta_j^{\sim}$ values then form an exponential model in (3.22) with selected $\theta_j^{\sim}$.

With stress data, $\theta_3^{\sim} = -0.2591$, $\theta_2^{\sim} = 0.2456$, $\theta_1^{\sim} = -0.2367$ were selected. And exponential model was formed in equation (3.23).

Step 4: Estimate coefficients of the exponential model.

Coefficients of the exponential model are estimated through Newton-Raphson iteration scheme in section 3.2.4.2. Then the exponential model formed with estimated components can be written as

$$d^{\wedge}(u) = \exp(\theta_0^{\wedge} + \sum_{j\in\mathcal{C}} \theta_j^{\wedge}\psi_j(H_N^{-1}(u))) \tag{4.3}$$

where $\mathcal{C}$ is an index set of selected order. With stress data, exponential model with estimated components is

$$d^{\wedge}(u;\boldsymbol{\theta}) = \exp\big(-0.052-0.2171\psi_1\big(H^{-1}(u)\big)+0.0922\psi_2\big(H^{-1}(u)\big)-0.1862\psi_3\big(H^{-1}(u)\big)\big). \tag{4.4}$$

given in equation (3.40). See Figure 9.

Step 5: Check the goodness of fit of estimated comparison density function

Check for goodness of fit is done using definition of comparison distribution. By integrating estimated comparison density function $d^{\wedge}(u;H,F)$, we can compute $D^{\wedge}(u;H,F)$, smooth comparison distribution function . Regarding stress data, see Figure 11. Also, we add 95% bootstrap confidence interval to the plot of $d^{\wedge}(u)$ for better interpretation. The confidence interval is computed through percentile method with 500 bootstrap samples. See Figure 10.

4.3. Summary and Discussion: Stress Data

This data set was analyzed by Giampaoli and Singer (2004). Assuming normality and homoscedasticity,

- The two-sample t-test with $d.f. = 20$ yields a p-value$= 0.0595$
- The Wilcoxon rank-sum test yields a p-value$= 0.2929$

According to these statistics, there is no sufficient evidence for rejection of the null: mean blood pressure of subjects are the same under normal or stress conditions. Usually most people stop their data analysis at this point. However, through our two-sample data analysis procedure, we are able to find more features of the stress

data. The control group is more likely to have lower blood pressure level than the stress group does. This means that stress could have an effect on the level of blood pressure. This finding is the opposite of the t-test or the Wilcoxon rank-sum test results. And we could estimate smooth compatison distibution function with even small sample sizes. Also, since we select three components(order 1,2, and 3), there might be differences in the direction of the 1st,2nd and 3rd order score functions.

Table 8

Score function value up to order 4 with stress data

X	$\psi_1(H_N^{-1}(u))$	$\psi_2(H_N^{-1}(u))$	$\psi_3(H_N^{-1}(u))$	$\psi_4(H_N^{-1}(u))$
81.5	-1.6569	1.9515	-2.0031	1.9181
81.7	-1.4991	1.3952	-0.8530	0.0969
85.5	-1.3413	0.8945	0.0089	-0.9090
87.1	-1.1835	0.4494	0.6129	-1.2992
88.9	-1.0257	0.0600	0.9896	-1.2511
89.4	-0.8679	-0.2739	1.1691	-0.9202
89.6	-0.7101	-0.5522	1.1820	-0.4396
89.9	-0.5523	-0.7749	1.0586	0.0797
92.2	-0.2367	-1.0534	0.5241	0.9006
92.4	0.0789	-1.1096	-0.1912	1.0966
92.7	0.2367	-1.0543	-0.5408	0.9177
93.5	0.3945	-0.9434	-0.8445	0.5758
94.6	0.5523	-0.7768	-1.0719	0.1147
95.0	0.7101	-0.5547	-1.1927	-0.3995
95.4	0.8679	-0.2770	-1.1766	-0.8788
95.5	1.0257	0.0563	-0.9930	-1.2129
96.4	1.1835	0.4452	-0.6118	-1.2694
96.8	1.3413	0.8897	-0.0024	-0.8939
97.0	1.4991	1.3898	0.8654	0.0903
109.2	1.6569	1.9455	2.0221	1.8820

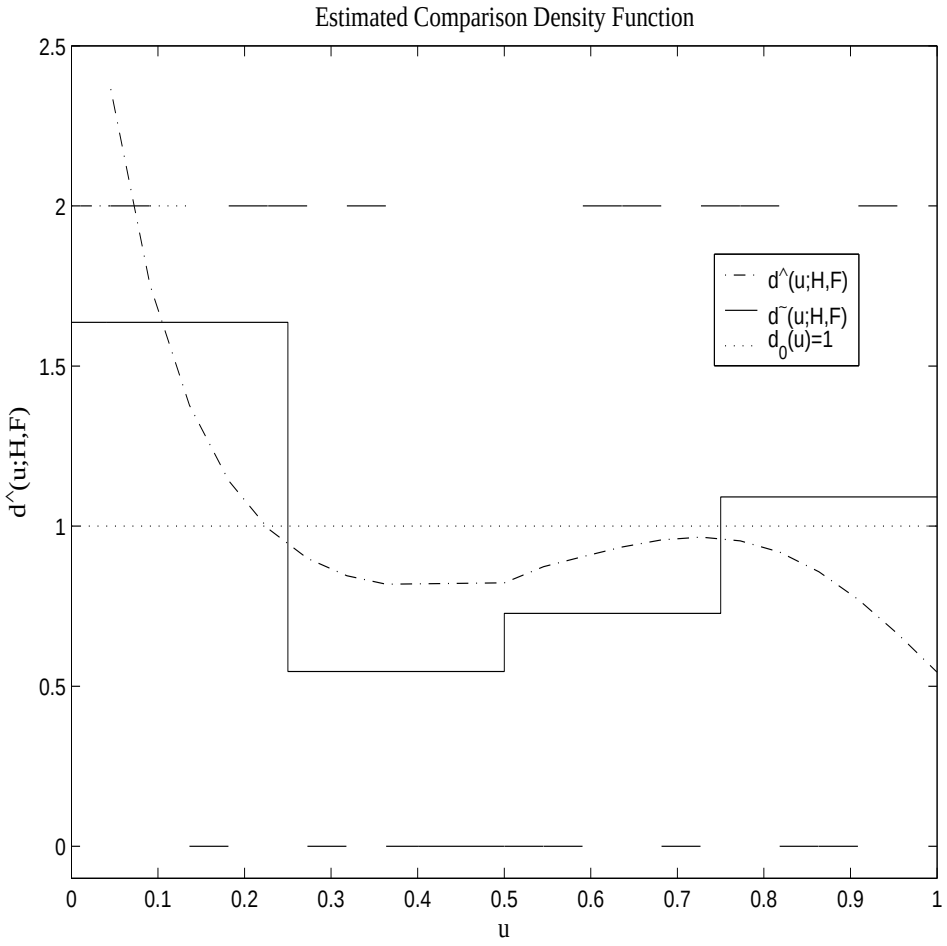

Figure 9. $d^{\wedge}(u; H, F)$: Estimated comparison density function through exponential model approach with stress data. The step function, the quartile density is added to the graph to see how exponential model approach works. The quartile density is defined for $i = 1, 2, 3, 4$ by $dQ_k(u) = 4\{D^{\sim}\big(i(.25)\big) - D^{\sim}\big((i-1)(.25)\big)\}$, $(i-1).25 < u < i(.25)$. For lower value of blood pressure($u < 0.25$), the comparison density is greater than 1, indicating a great frequency of observations in control group.

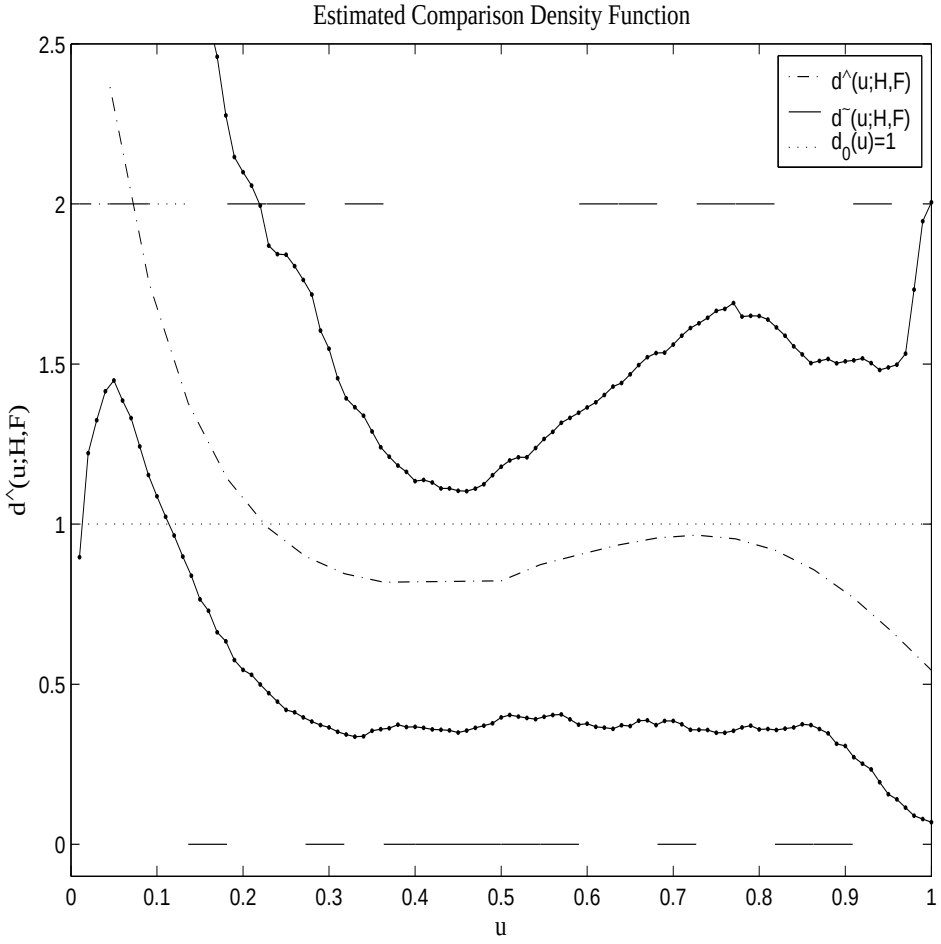

Figure 10. 95% bootstrap confidence interval of $d^{\wedge}(u; H, F)$: For better interpretation, bootstrap confidence interval is added and the confidence interval is computed through percentile method with 500 bootstrap samples. Since the confidence interval does not include uniform density $d_0(u)$, we conclude that the distributions of the blood pressure level of two groups are different and stress does have an effect on blood pressure level.

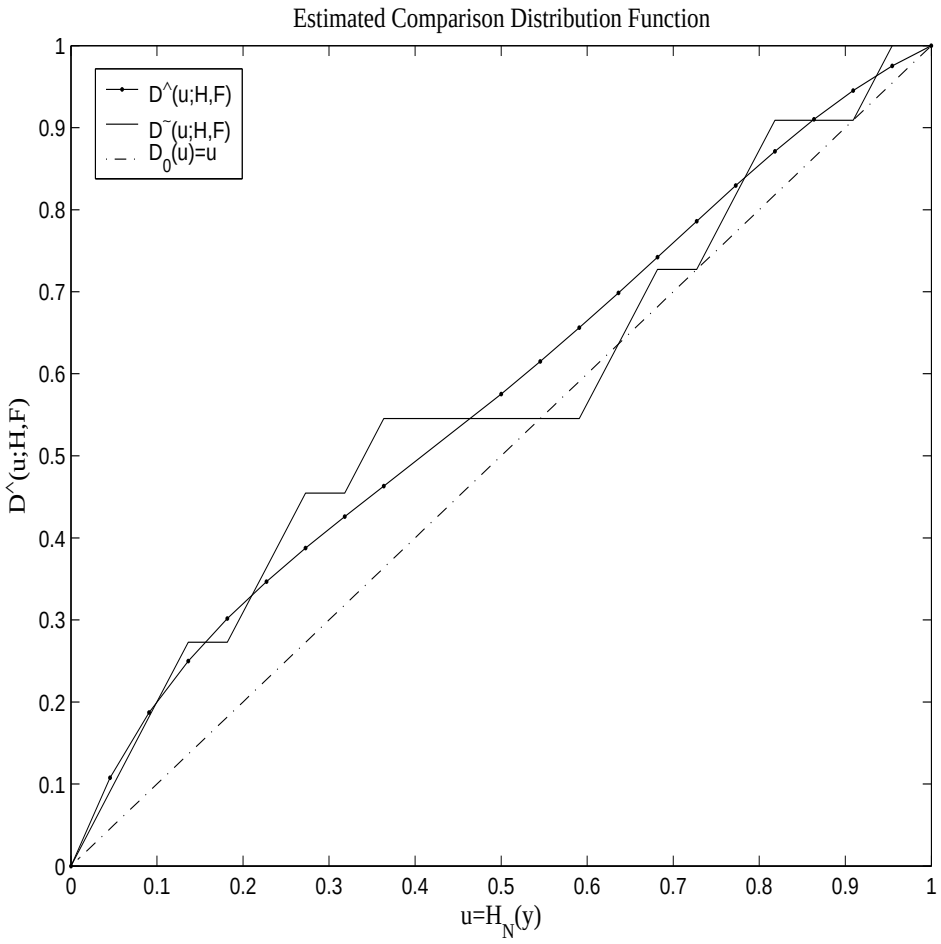

Figure 11. $D^{\wedge}(u; H, F)$: Estimated comparison distribution function with stress data. Since estimated comparison distribution function goes with $D^{\sim}(u; H, F)$ very well, we conclude that our exponential model estimation is working properly.

CHAPTER V

EXAMPLES AND APPLICATIONS

5.1. Introduction

The two-sample data analysis procedure derived in the Chapter IV is applied to another example of real data. In the section 5.2, we provide basic information related with the data set. In the section 5.3 before performing our analysis procedure, we have a summary of explanatory analysis of the data which was done by Parzen (2004). Then data analysis through exponential model approach will be performed in the section 5.4.

5.2. Radon Cancer Data

To illustrate our new procedure of the two-sample data analysis, we consider a data set from an article "Indoor radon and childhood cancer"(Wakefield and Kohler (1991)) which we call as radon cancer data. To study the effect of indoor radon concentration to incidence of childhood cancer, a case-control study was done measuring indoor radon concentrations over the same 3-month period in bedroom and living room of children in the Wessex health region. The cases were composed of children with cancer diagnosed within the previous 3 years and controls were matched for age and area of residence. For the data set, see Table 9. The data have two independent samples from cancer group and no-cancer group and researchers want to know how indoor radon effects the incidence of childhood cancer. This is one of the typical two-sample problems. Instead of testing the homogeneity of locations or scales, we will

Table 9

Radon concentration levels

Cancer	Cancer	Non-cancer	Non-cancer
3	16	3(2)	12
5	16	3	12
6	17	5	13
7	18	6	14
8	18	6	17
9	18	7	17
9	20	7	21
10	21	7	21
10	21	8	24
10	22	8	24
11	22	8	29
11	23	9	29
11	23	9	29
11	27	9	29
12	33	9	33
13	34	11	39
13	38	11	55
15	39	11	55
15	45	11	85
15	57	11	
16			

extract more information from the data by testing $H_0 : F = H$ where F is continuous distribution for the first group and H is pooled distribution. Let $X_1, \cdots, X_{39} (m = 39)$ be a sample from no-cancer group with distribution function F, and $Y_1, \cdots, Y_{41} (n = 41)$ be a sample from cancer group with distribution function G and thus $N = m+n = 80$. Thus $\lambda = m/N = 39/80 = 0.4875$.

5.3. Explanatory Data Analysis

Explanatory analysis of radon cancer data was performed by Parzen (2004). Parzen provided the following conclusions on radon cancer data;

- Location parameter is greater in cancer houses than in non-cancer houses.
- Variability of radon in non-cancer houses is greater than that of radon in cancer houses.
- Non-cancer homes radon has skew distribution and cancer homes radon has symmetric distribution.
- Non-cancer houses radon is fitted by exponential distribution and cancer houses radon is fitted by normal distribution with outliers.

Also, Parzen presented P-P plot of two sample distributions which estimate the pooled comparison distribution $D(u; H, F)$. For the related plots and tables, see Parzen (2004).

5.4. Two-sample Data Analysis Using Exponential Model Approach

In this section, our two-sample data analysis algorithm is applied to the radon cancer data. For each step, we have corresponding interpretations too.

- Step 1: Combine two samples and arrange them in order. Estimate comparison distribution $D(u; H, F)$ by drawing a P-P plot. H-exact values for distinct y_j values is given in the Table 10. Figure 12 is the corresponding P-P plot. Even though P-P plot seems to be close to the 45 degree straight line, since $D^{\sim}(u; H, F)$ is usually very rough, we proceed to step 2.

- Step 2: Compute mid-distribution score functions ψ_j up to order 4. To verify the orthonormality of computed score functions, see Table 11 and Figure 13.

- Step 3: Compute the estimated values of components $\theta_j^{\sim}$.

$$\begin{aligned} \theta_1^{\sim} &= -0.1268, \\ \theta_2^{\sim} &= 0.2074, \\ \theta_3^{\sim} &= 0.0970, \\ \theta_4^{\sim} &= -0.0960. \end{aligned}$$

  We select the first and the second components to form an exponential model. Also, a quadratic pattern from the Figure 14 supports our components selection.

- Step 4: Estimate coefficients of the exponential model through Newton-Raphson iteration. Exponential model with estimated components is

$$d^{\wedge}(u;\boldsymbol{\theta}) = \exp\Big(-0.0276 - 0.1841\psi_1\big(H^{-1}(u)\big) + 0.1298\psi_2\big(H^{-1}(u)\big)\Big). \quad (5.1)$$

  given in the equation (3.40). See Table 12 and Figure 15.

- Step 5: Check the goodness of fit of estimated comparison density function using definition of comparison distribution. By integrating estimated comparison density function $d^{\wedge}(u;H,F)$, we can compute $D^{\wedge}(u;H,F)$, smooth comparison distribution function . With stress data, see Figure 16. 95% bootstrap confidence interval is added to the plot of $d^{\wedge}(u)$ for better interpretation. The confidence interval is computed through percentile method with 500 bootstrap

samples. See Figure 17.

5.5. Summary and Discussion: Radon Cancer Data

This data set was analyzed by Wakefield and Kohler (1991) and Parzen (2004) from quite different views. Wakefield and Kohler (1991) concluded that there was no significant difference between the mean indoor radon concentration levels for no-cancer group and cancer group. However, Parzen (2004) pointed out the differences in distributions between two groups as well as those in location and variability. Through our two-sample data analysis procedure, we try to find more features of radon cancer data. From Figure 15, we clearly see that no-cancer group is more likely to have lower indoor radon concentration level than cancer group does. This means that indoor radon concentration level could have an effect on incidence of childhood cancers. And we could have this finding with even small sample sizes. Also, since we select two components(order 1 and 2), there might be differences in direction of 1st and 2nd order score functions which indicates differences in location and scale.

Table 10

Sample pooled distribution function H_N. The number in () means the number of occurrences of the corresponding observation.

Radon concentration	H_N
3.00(3)	0.0375
5.00(2)	0.0625
6.00(3)	0.1000
7.00(4)	0.1500
8.00(4)	0.2000
9.00(6)	0.2750
10.00(3)	0.3125
11.00(9)	0.4250
12.00(3)	0.4625
13.00(3)	0.5000
14.00	0.5125
15.00(3)	0.5500
16.00(3)	0.5875
17.00(3)	0.6250
18.00(3)	0.6625
20.00	0.6750
21.00(4)	0.7250
22.00(2)	0.7500
23.00(2)	0.7750
24.00(2)	0.8000
27.00	0.8125
29.00(4)	0.8625
33.00(2)	0.8875
34.00	0.9000
38.00	0.9125
39.00(2)	0.9375
45.00	0.9500
55.00(2)	0.9750
57.00	0.9875
85.00	1.0000

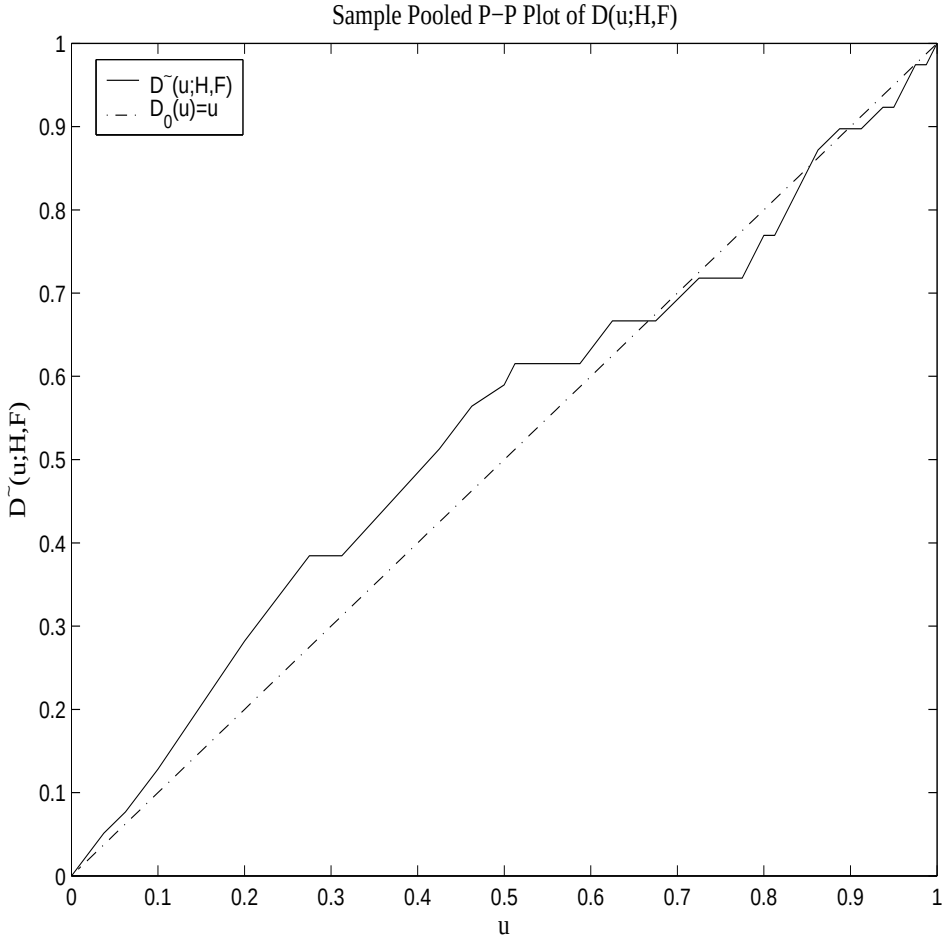

Figure 12. Sample pooled comparison distribution function with radon cancer data.

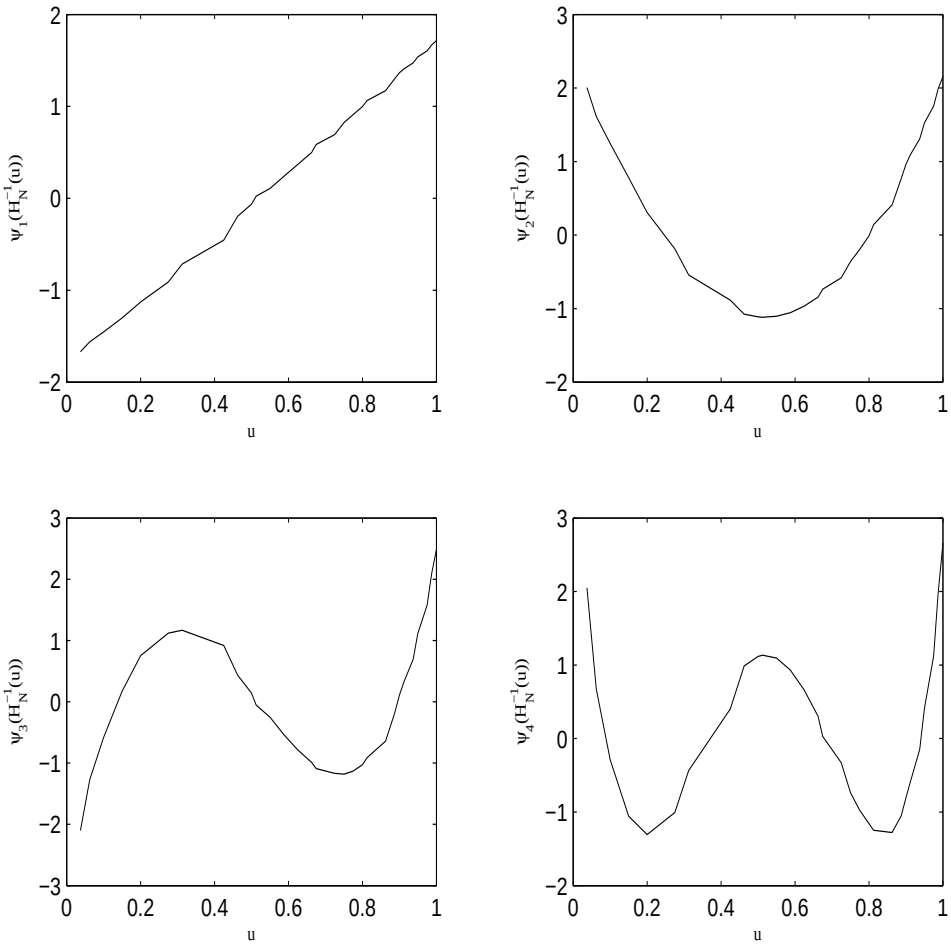

Figure 13. Sample mid-distribution score functions up to order 4 using radon cancer data. $\psi_j\big(H_N^{-1}\big(u\big)\big)$ for $u_{i-1} < u < u_i$ and $H_N^{-1}(u_i) = z_i$.

Table 11

Inner product of score functions to verify orthonormality with radon cancer data

	$\psi_1(H_N^{-1}(u))$	$\psi_2(H_N^{-1}(u))$	$\psi_3(H_N^{-1}(u))$	$\psi_4(H_N^{-1}(u))$
$\psi_1(H_N^{-1}(u))$	1	7.6328e-017	-1.2698e-015	3.8858e-016
$\psi_2(H_N^{-1}(u))$	7.6328e-017	1	-4.0246e-016	-3.1919e-016
$\psi_3(H_N^{-1}(u))$	-1.2698e-015	-4.0246e-016	1	2.4564e-015
$\psi_4(H_N^{-1}(u))$	3.8858e-016	-3.1919e-016	2.4564e-015	1

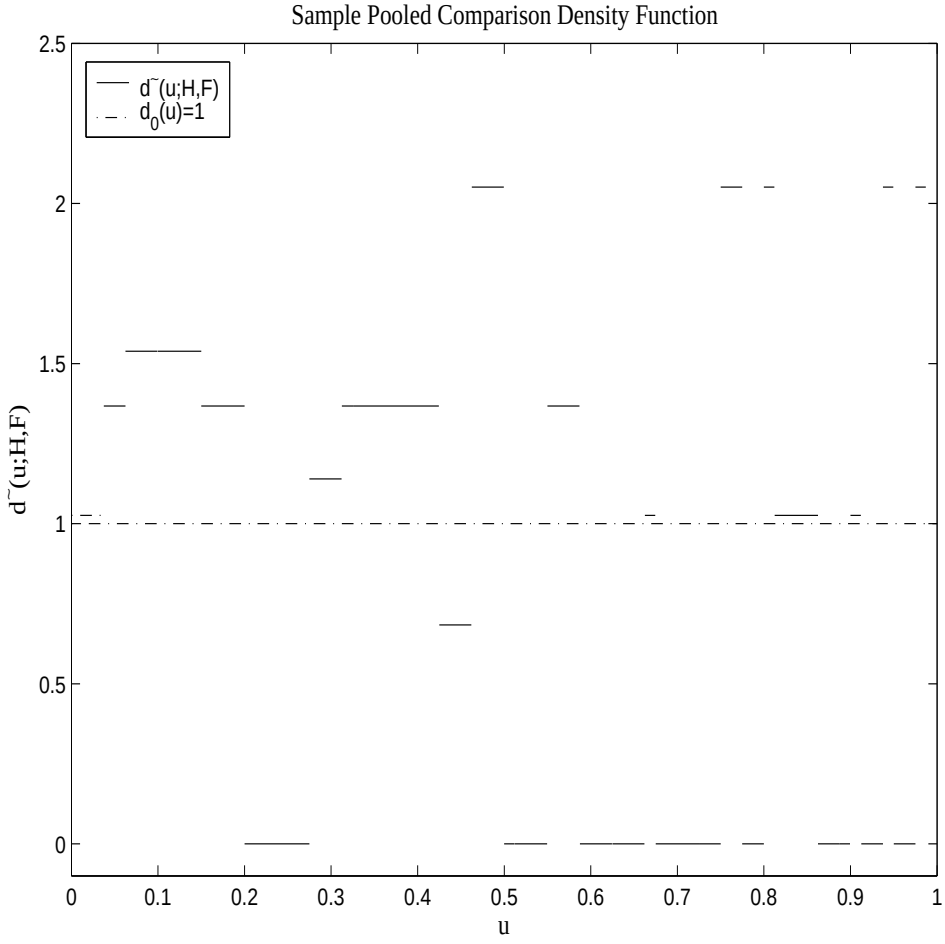

Figure 14. $d^{\sim}(u; H.F)$: Sample comparison density function with radon cancer data.

Table 12

$\theta_j^{\wedge}$ values up to order 2 through Newton-Raphson iteration with radon cancer data

Iteration	θ_0	θ_1	θ_2
1	-0.0320	-0.1268	0.2074
2	-0.0274	-0.1784	0.1366
3	-0.0276	-0.1841	0.1298
4	-0.0276	-0.1841	0.1298

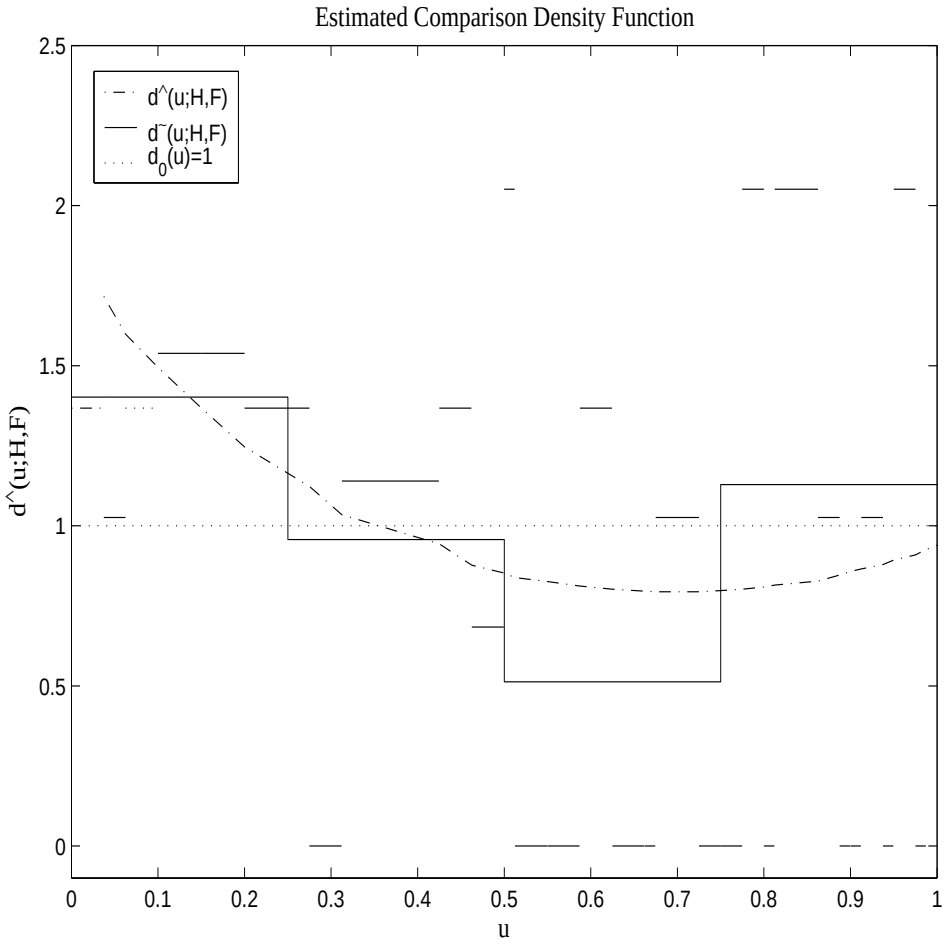

Figure 15. $d^{\wedge}(u; H, F)$: Estimated comparison density function through exponential model approach with two components with radon cancer data.

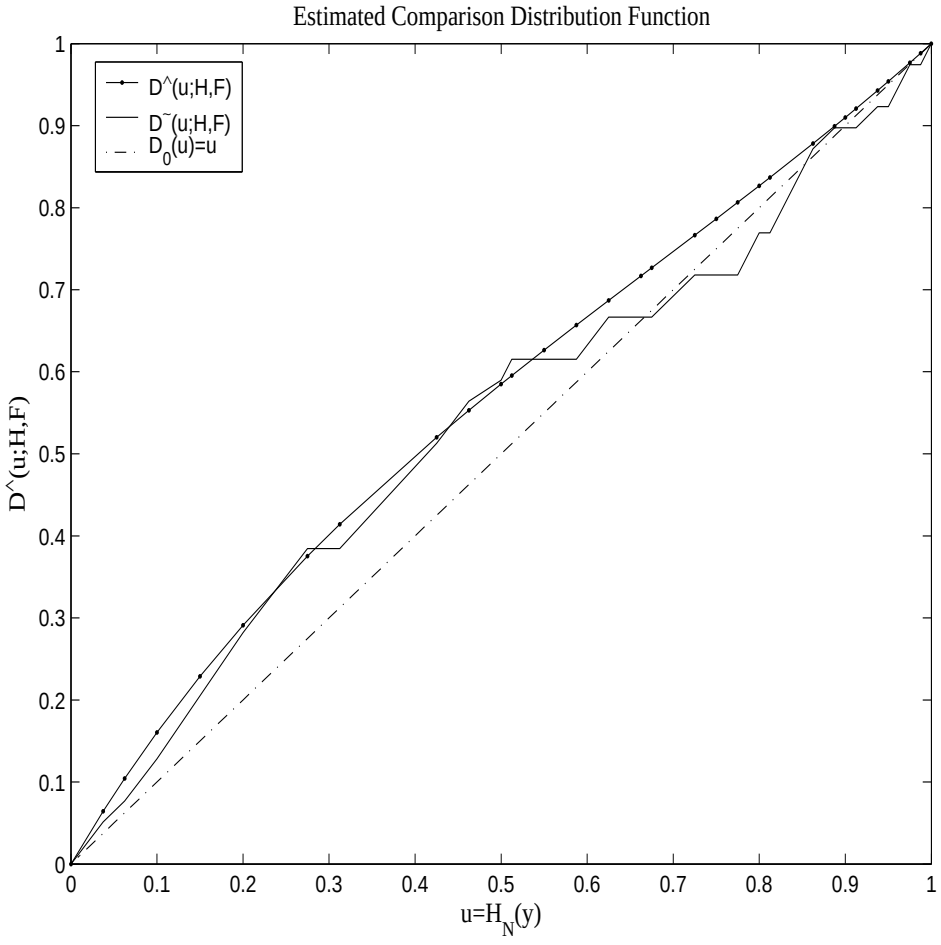

Figure 16. $D^{\wedge}(u; H, F)$: Estimated comparison distribution function with radon cancer data with 2 components. Since estimated comparison distribution function goes with $D^{\sim}(u; H, F)$ very well, we conclude that our exponential model estimation is working properly.

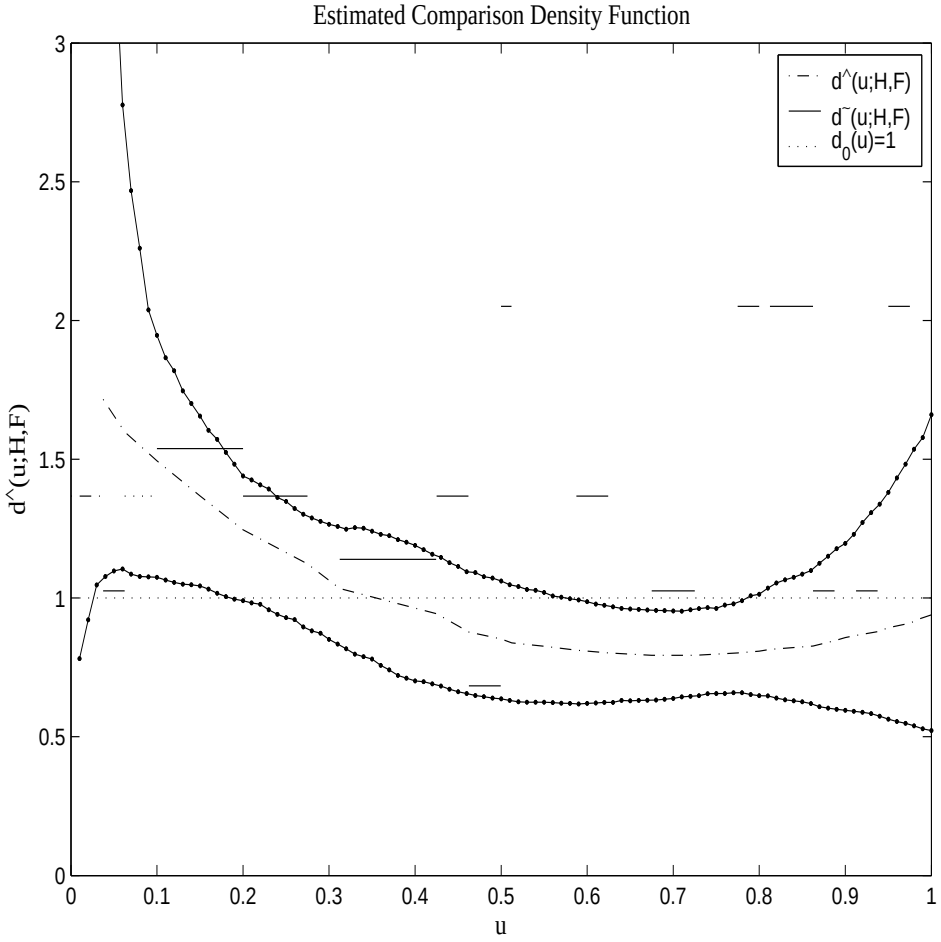

Figure 17. 95% bootstrap confidence interval of $d^{\wedge}(u; H, F)$: For better interpretation, bootstrap confidence interval is added and the confidence interval is computed through percentile method with 500 bootstrap samples. Since the confidence interval does not include uniform density $d_0(u)$, we conclude that the distributions of the radon concentration level of two groups are different and radon does have an effect on childhood cancer incidence.

5.6. Simulation Results

We apply our two-sample data analysis procedure to several cases to see how comparison distribution function and comparison density function behave and to find structured interpretation rules. We have all 8 possible cases according to differences in either locations or scales or distributions. For each case, we have probability functions, estimated comparison density functions and comparison distribution functions. Let X be a random variable. Then $X \sim Normal(\mu, \sigma^2)$ means that a random variable X has a normal distribution with a density function

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(X-\mu)^2}{2\sigma^2}\right). \tag{5.2}$$

where $-\infty < x < \infty$, $E(X) = \mu$, and $Var(X) = \sigma^2$. In the same way, $X \sim Gamma(\lambda, \gamma)$ means that a random variable X has a gamma distribution with a density function

$$f(x) = \frac{\lambda^\gamma}{\Gamma(\gamma)} x^{\gamma-1} \exp(-\lambda x) \tag{5.3}$$

where $x > 0$, $\lambda > 0$ and $\gamma > 0$. And $E(X) = \gamma/\lambda$, and $Var(X) = \gamma/\lambda^2$. We generate two-sample data using these two distributions and apply our exponential model approach to the generated data set. From each distribution, 100 samples are generated. To represent two samples, we use two random variables X and Y. As a measure of location and scale, we use mean($E(X)$) and variance($Var(X)$) respectively. For corresponding example for each case, see Table 13.

Table 13

All possible cases according to differences in either locations or scales or distributions. "0" means that there are no differences between two samples and "1" means there are differences between two samples

Case number	Locations	Scales	Distributions
1	0	0	0
2	0	0	1
3	0	1	0
4	1	0	0
5	0	1	1
6	1	0	1
7	1	1	0
8	1	1	1

5.6.1. Case 1: Same Distributions, Same Locations, and Same Scales

In this case, we know that comparison density function is uniform distribution. Thus, we omit the simulation result.

5.6.2. Case 2: Same Locations, Scales but Different Distributions

Let $X \sim N(1, 1^2)$ and $Gamma(1, 1)$ with $E(X) = E(Y) = 1$ and $Var(X) = Var(Y) = 1$ under different distributions. Figure 18, Figure 19 and Figure 20 are plots of two density functions, $d^{\wedge}(u)$ and $D^{\wedge}(u)$ respectively.

5.6.3. Case 3: Same Locations, Different Scales and Same Distributions

Let $X \sim Normal(0, 5^2)$ and $Y \sim Normal(0, 1^2)$. Then $E(X) = E(Y) = 0$ and $Var(X) = 5^2$ and $Var(Y) = 1^2$ under normal distribution. Figure 21, Figure 22 and Figure 23 are plots of two density functions, $d^{\wedge}(u)$ and $D^{\wedge}(u)$ respectively.

5.6.4. Case 4: Different Locations, but Same Scales and Distributions

Let $X \sim Normal(0, 1^2)$ and $Y \sim Normal(3, 1^2)$. Then $E(X) = 0$ and $E(Y) = 3$ and $Var(X) = Var(Y) = 1^2$ under normal distribution. Figure 24, Figure 25 and Figure 26 are plots of two density functions, $d^{\wedge}(u)$ and $D^{\wedge}(u)$ respectively.

5.6.5. Case 5: Same Locations, but Different Scales and Distributions

Let $X \sim N(2, 1^2)$ and $Y \sim Gamma(1, 2)$ with $E(X) = E(Y) = 2$ and $Var(X) = 1^2$ and $Var(Y) = 2$ under different distributions. We select 2nd,3rd and 4th order to form an exponential model. Figure 27, Figure 28 and Figure 29 are plots of two density functions, $d^{\wedge}(u)$ and $D^{\wedge}(u)$ respectively.

5.6.6. Case 6: Different Locations, Same Scales and Different Distributions

Let $X \sim Normal(0, \sqrt{2}^2)$ and $Y \sim Gamma(1, 2)$ with $E(X) = 0$, $E(Y) = 2$ and $Var(X) = Var(Y) = 2$ under different distributions. Figure 30, Figure 31 and Figure 32 are plots of two density functions, $d^{\wedge}(u)$ and $D^{\wedge}(u)$ respectively.

5.6.7. Case 7: Different Locations, Scales but Same Distributions

Let $X \sim Normal(0, 1^2)$,$Y \sim Normal(3, 2^2)$. Then, $E(X) = 0$, $E(Y) = 3$ and $Var(X) = 1^2$ and $Var(Y) = 2^2$ under normal distribution. Figure 33, Figure 34 and Figure 35 are plots of two density functions, $d^{\wedge}(u)$ and $D^{\wedge}(u)$ respectively.

5.6.8. Case 8: Different Locations, Scales and Distributions

Let $X \sim Normal(0, 1^2)$ and $Y \sim Gamma(2/3, 2)$. Then, $E(X) = 0$, $E(Y) = 3$ and $Var(X) = 1$ and $Var(Y) = 4.5$ under different distributions. Figure 36, Figure 37 and Figure 38 are plots of two density functions, $d^{\wedge}(u)$ and $D^{\wedge}(u)$ respectively.

5.6.9. Summary and Discussion

As a measure of location and scale, we use mean and variance. And when we say the difference in distribution, usually that is about differences in skewness and excess. Thus we try to see how comparison density function and comparison distribution function behave according to these features.

Figure 22 shows a case of difference only in distribution. If there is difference only in location between two groups, estimated comparison density will show monotone linear pattern(Figure 25). And if there is difference only in scale between two groups, estimated comparison density will show symmetric quadratic pattern(Figure 25). Case 4 and case 6 show similar patterns(Figure 25 and Figure 31) and both cases

have differences in location and no differences in scale. That gives monotone linear pattern to each case. Also, from Figure 32, we see difference in distribution through asymmetry compared with Figure 26 which shows symmetry. Also, case 7 and case 8 show analogous patterns(Figure 34 and Figure 37) and both cases have differences in location and scale. If we see Figure 33 and Figure 36, we can also see similarity.

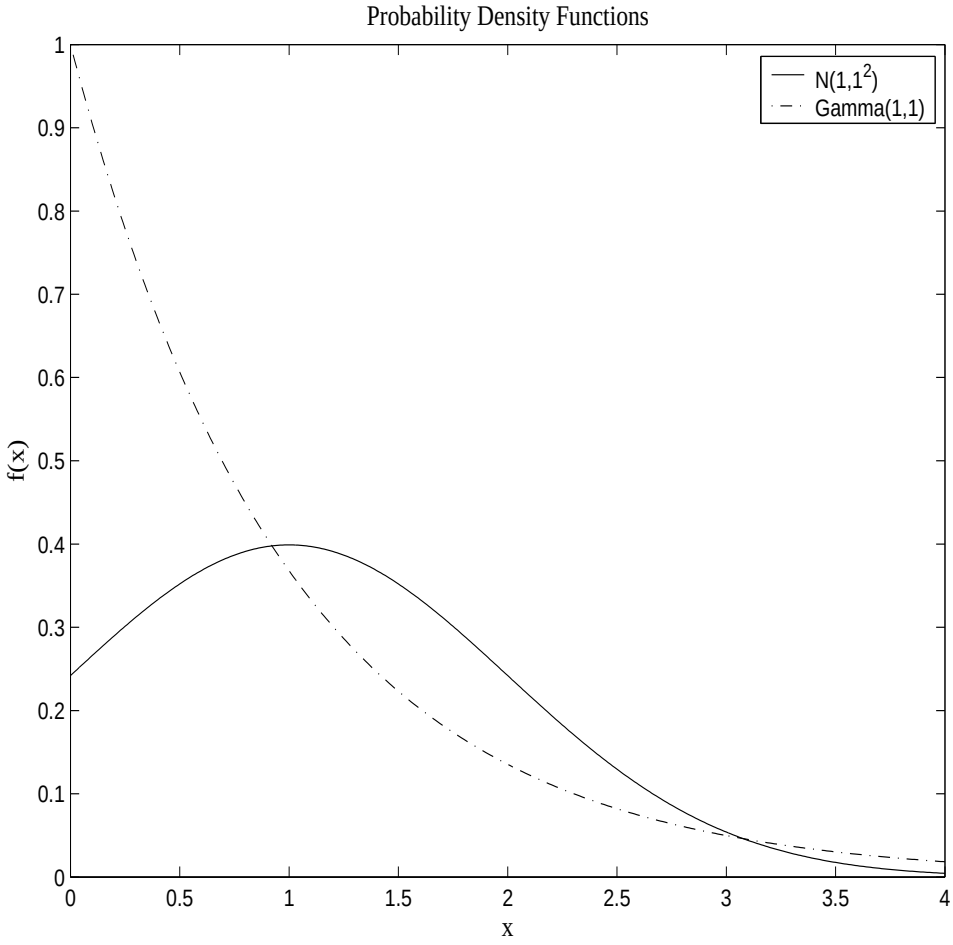

Figure 18. Case 2: Same locations, scales but different distributions: Probability density functions of $X \sim Normal(1, 1^2)$ and $Y \sim Gamma(1, 1)$.

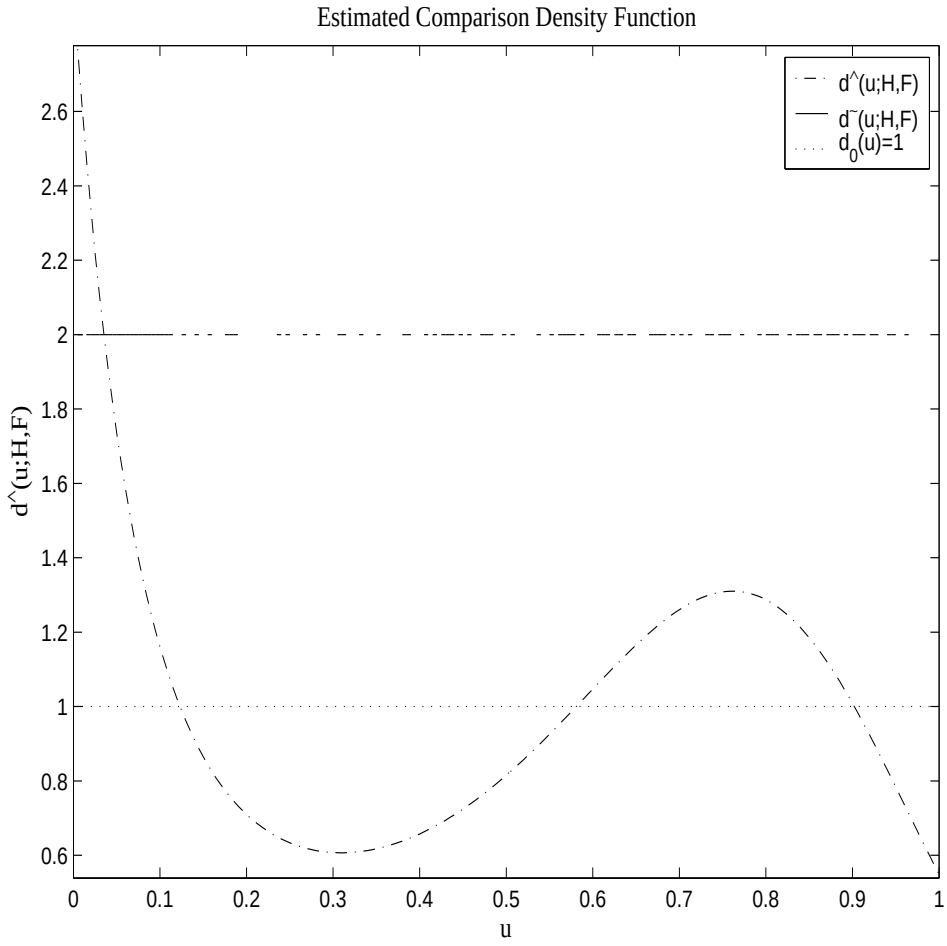

Figure 19. Case 2: Same locations, scales but different distributions: $d\hat{}(u; H, F)$: Estimated comparison density function with $X \sim Normal(1, 1^2)$ and $Y \sim Gamma(1, 1)$. 2nd and 3rd order score functions were selected($\mathcal{C} = \{2, 3\}$).

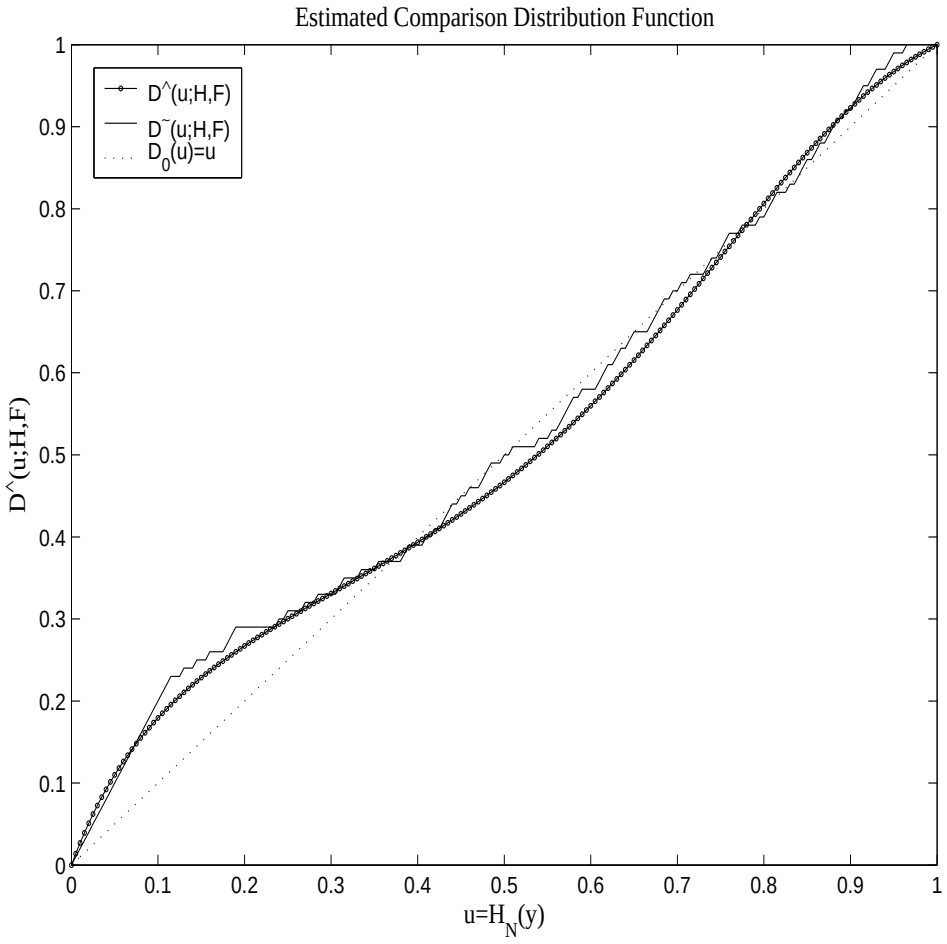

Figure 20. Case 2: Same locations, scales but different distributions: $D^{\wedge}(u; H, F)$: Estimated comparison distribution function with $X \sim Normal(1, 1^2)$ and $Y \sim Gamma(1, 1)$.

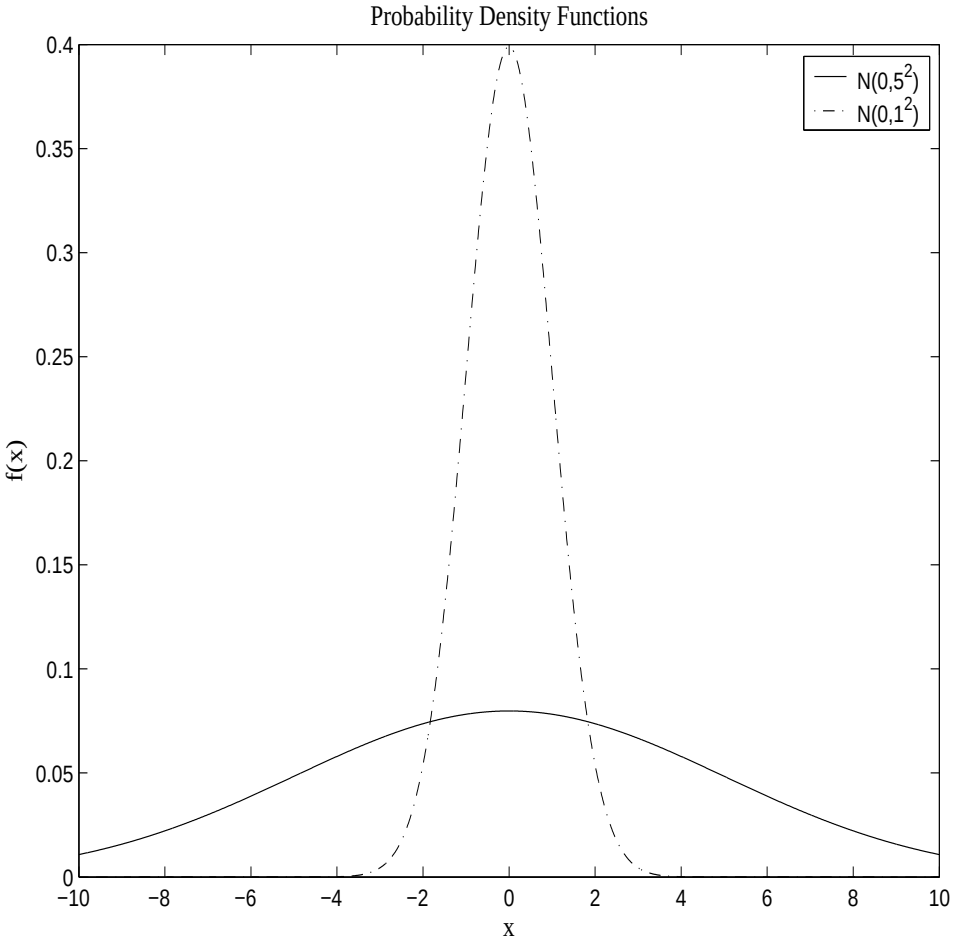

Figure 21. Case 3: Same locations, different scales and same distributions: Probability density functions of $X \sim Normal(0, 5^2)$ and $Y \sim Normal(0, 1^2)$.

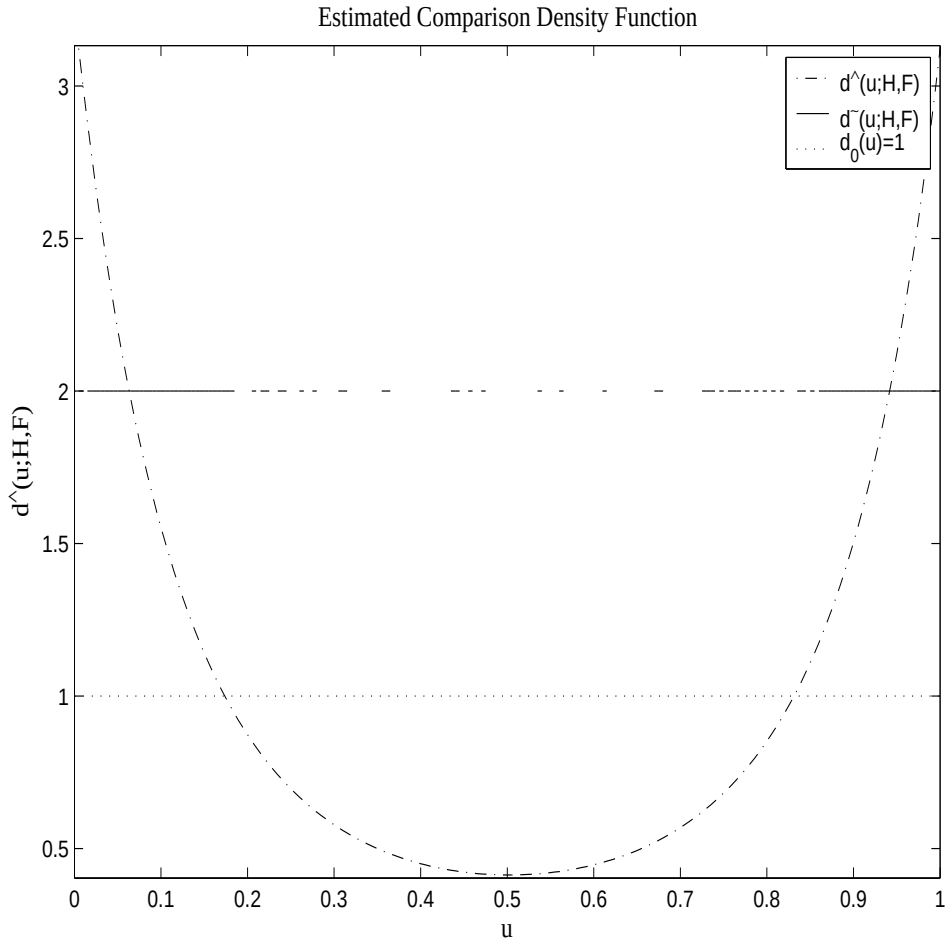

Figure 22. Case 3: Same locations, different scales and same distributions: $d^{\wedge}(u; H, F)$: Estimated comparison density function with $X \sim Normal(0, 5^2)$ and $Y \sim Normal(0, 1^2)$. Only 2nd order score function was selected($\mathcal{C} = \{2\}$).

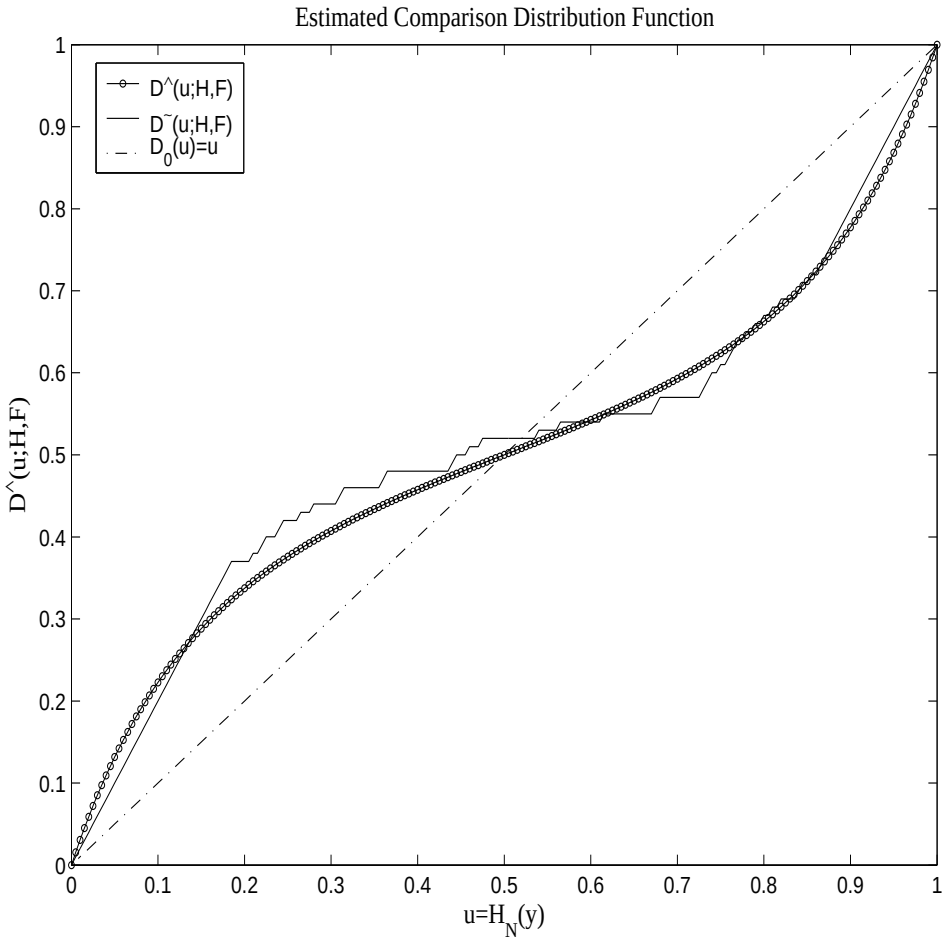

Figure 23. Case 3: Same locations, different scales and same distributions: $D^{\wedge}(u; H, F)$: Estimated comparison distribution function with $X \sim Normal(0, 5^2)$ and $Y \sim Normal(0, 1^2)$

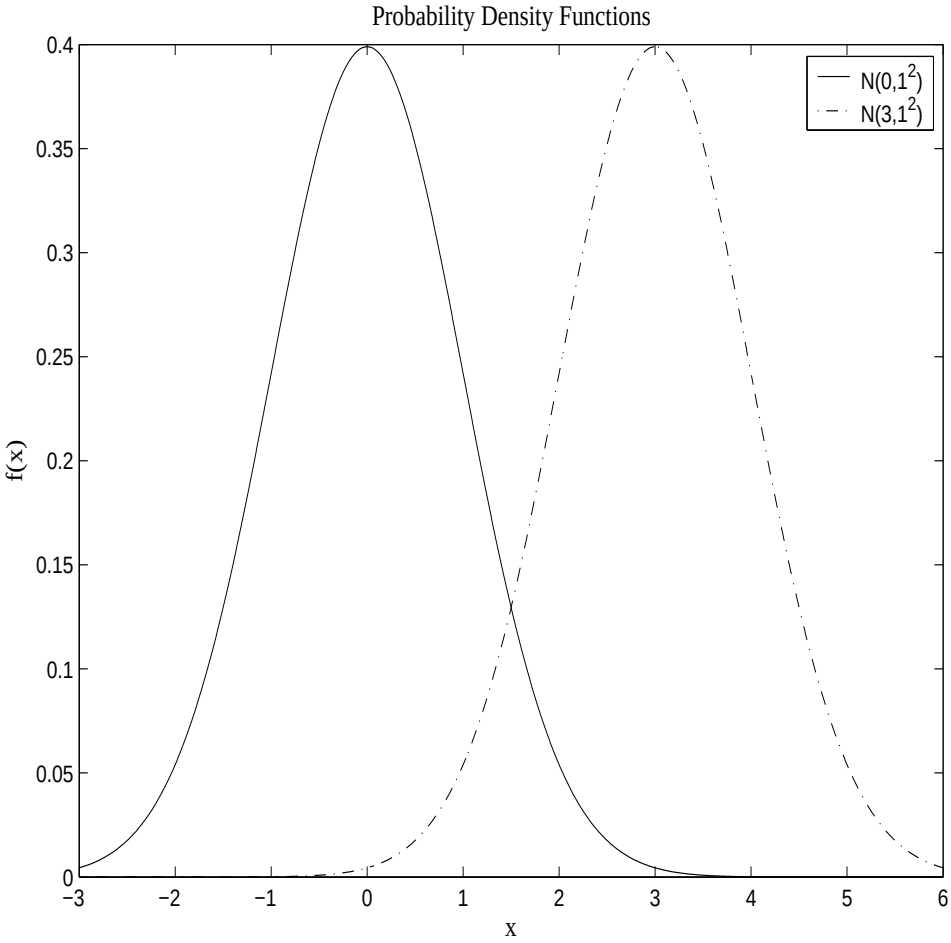

Figure 24. Case 4: Different locations, but same scales and distributions: Probability density functions of $X \sim Normal(0, 1^2)$ and $Y \sim Normal(3, 1^2)$.

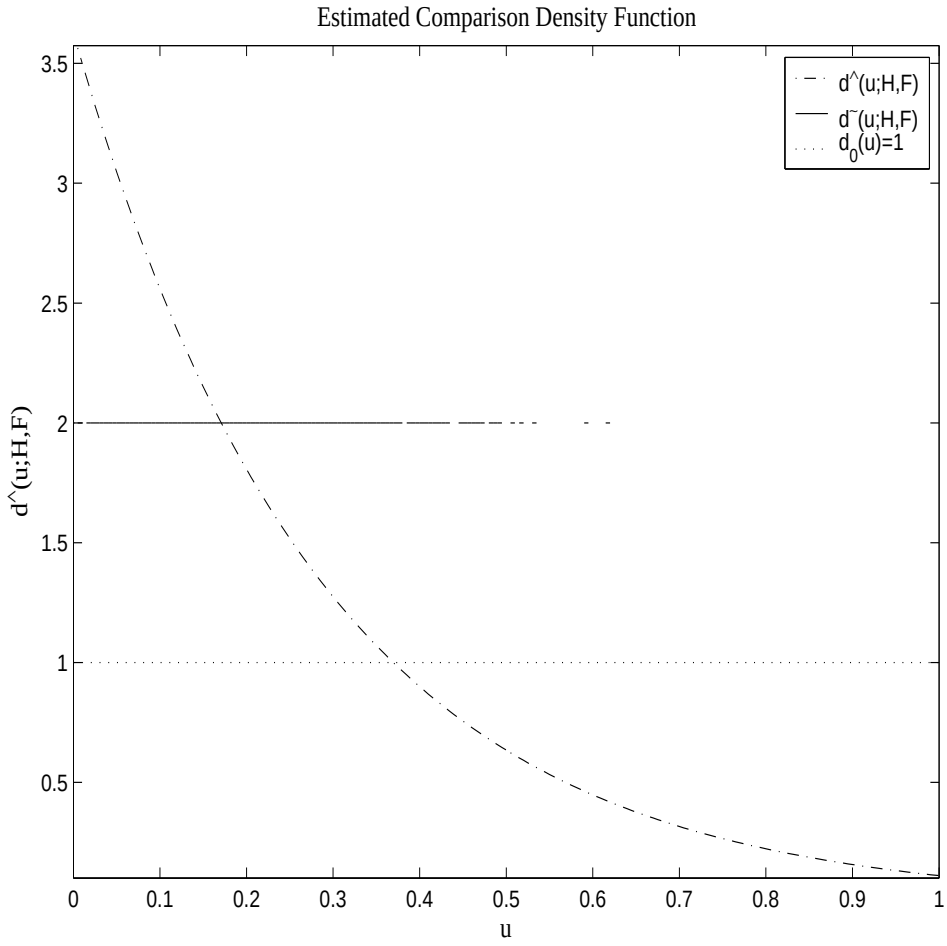

Figure 25. Case 4: Different locations, but same scales and distributions: $d^{\wedge}(u; H, F)$: Estimated comparison density function with $X \sim Normal(0, 1^2)$ and $Y \sim Normal(3, 1^2)$. Only 1st order component was selected($\mathcal{C} = \{1\}$).

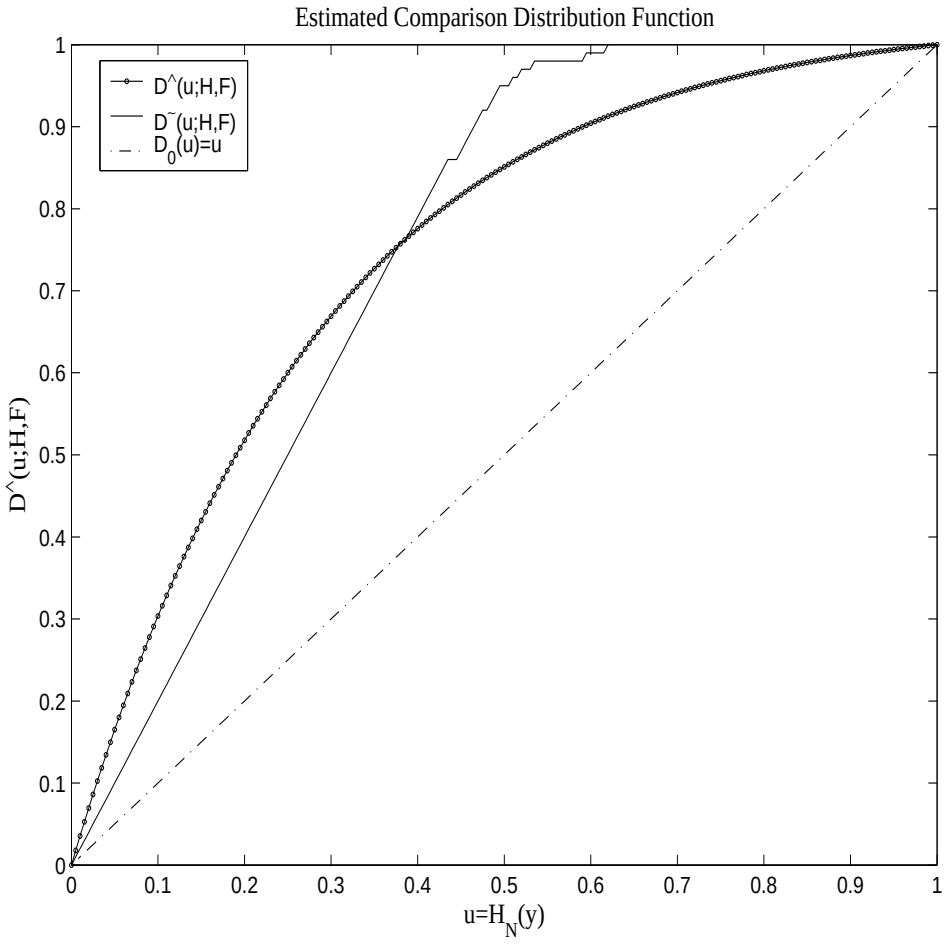

Figure 26. Case 4: Different locations, but same scales and distributions: $D^{\wedge}(u; H, F)$: Estimated comparison distribution function with $X \sim Normal(0, 1^2)$ and $Y \sim Normal(3, 1^2)$

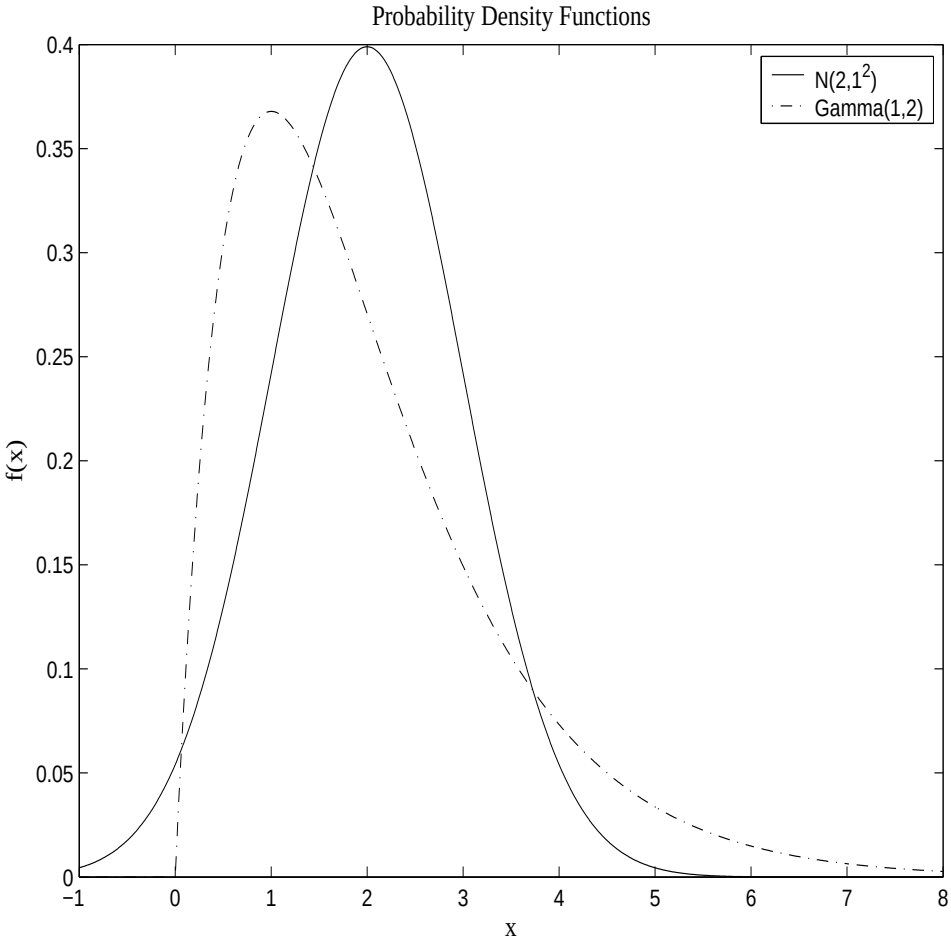

Figure 27. Case 5: Same locations, but different scales and distributions: Probability density functions of $X \sim Normal(2, 1^2)$ and $Y \sim Gamma(1, 2)$.

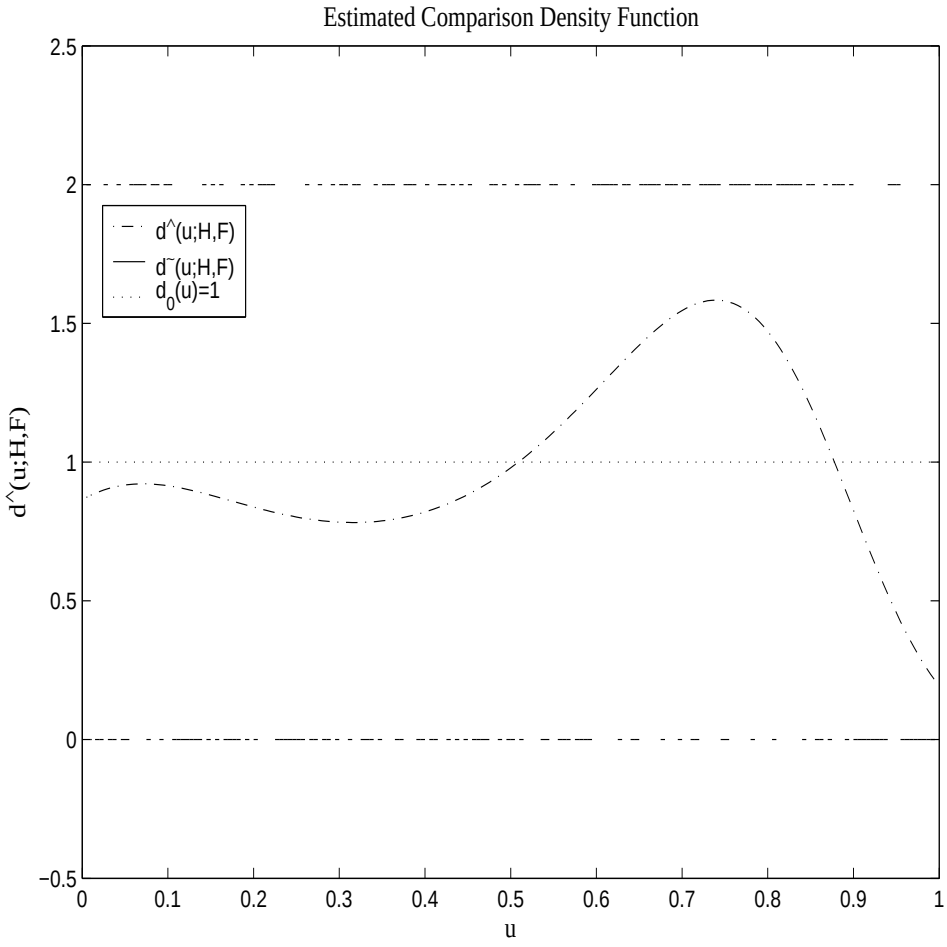

Figure 28. Case 5: Same locations, but different scales and distributions: $d^{\wedge}(u; H, F)$: Estimated comparison density function with $X \sim Normal(2, 1^2)$ and $Y \sim Gamma(1, 2)$. 2nd, 3rd, and 4th order score functions were selected($\mathcal{C} = \{2, 3, 4\}$).

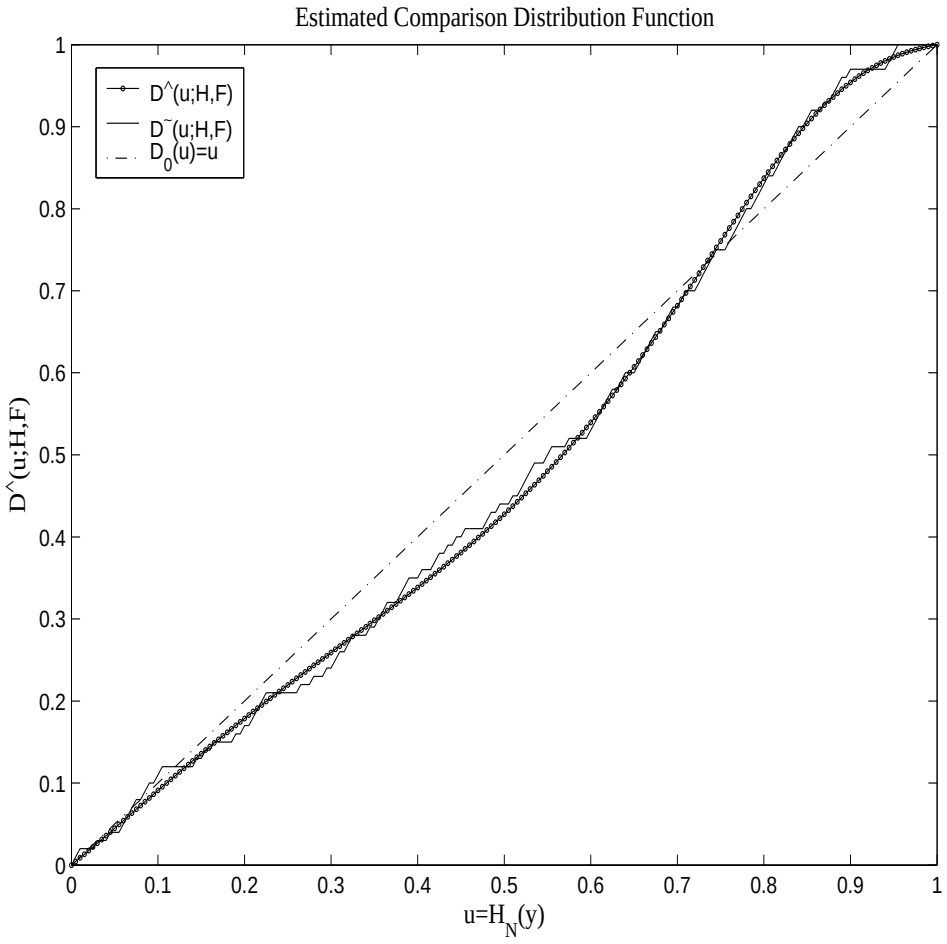

Figure 29. Case 5: Same locations, but different scales and distributions: $D\hat{}(u; H, F)$: Estimated comparison distribution function with $X \sim Normal(0, 1^2)$ and $Y \sim Normal(3, 1^2)$

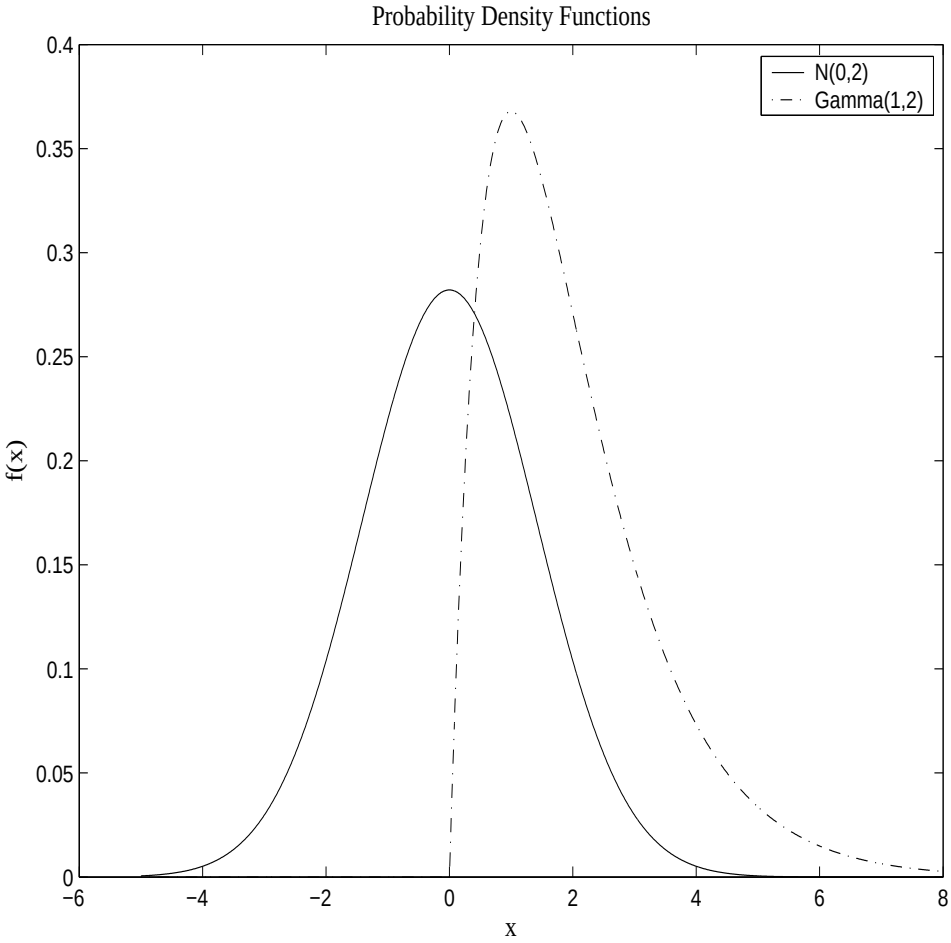

Figure 30. Case 6: Different locations, same scales and different distributions: Probability density functions of $X \sim Normal(0, \sqrt{2}^2)$ and $Y \sim Gamma(1, 2)$.

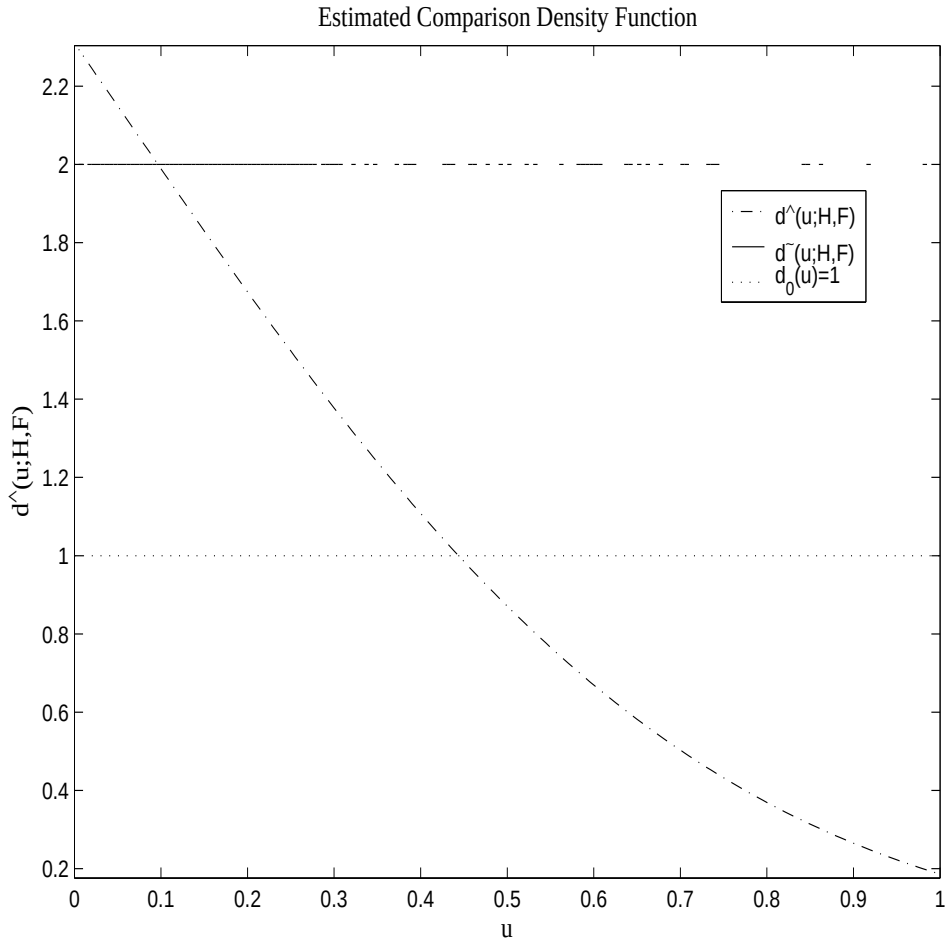

Figure 31. Case 6: Different locations, same scales and different distributions: $d^{\wedge}(u; H, F)$: Estimated comparison density function with $X \sim Normal(0, \sqrt{2}^2)$ and $Y \sim Gamma(1, 2)$. 1st and 2nd order score functions were selected($\mathcal{C} = \{1, 2\}$).

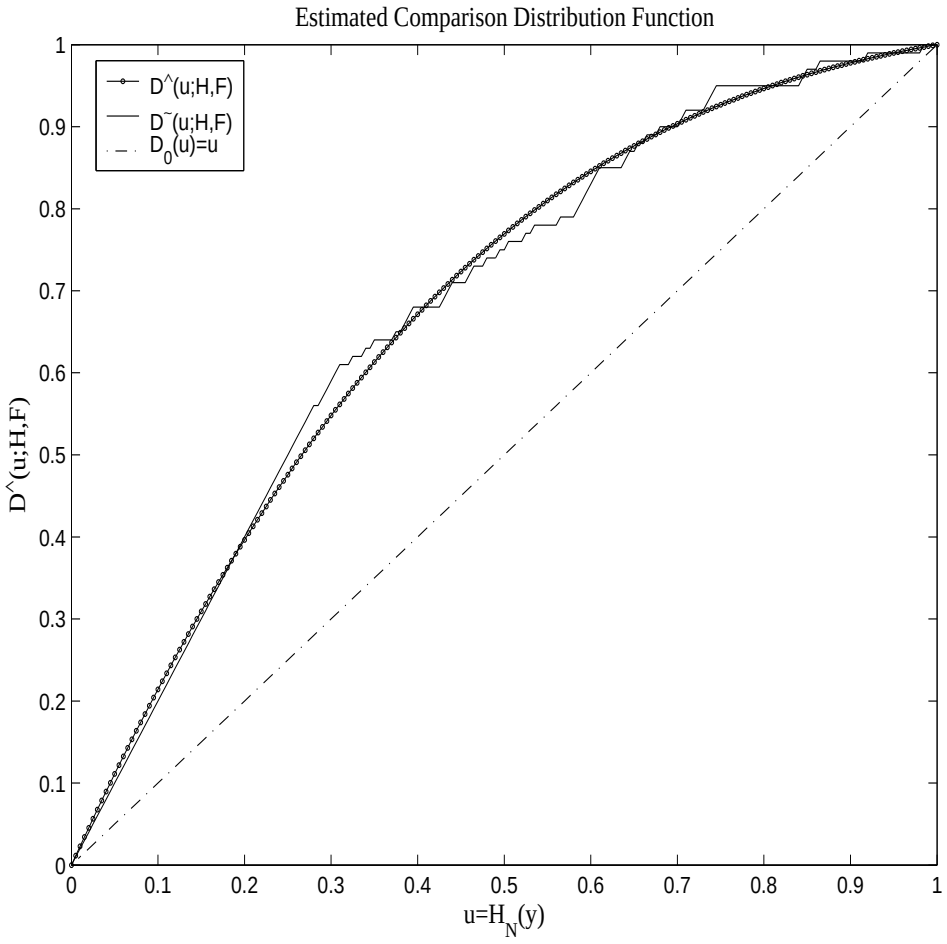

Figure 32. Case 6: Different locations, same scales and different distributions: $D^{\wedge}(u; H, F)$: Estimated comparison distribution function with $X \sim Normal(0, \sqrt{2}^2)$ and $Y \sim Gamma(1, 2)$

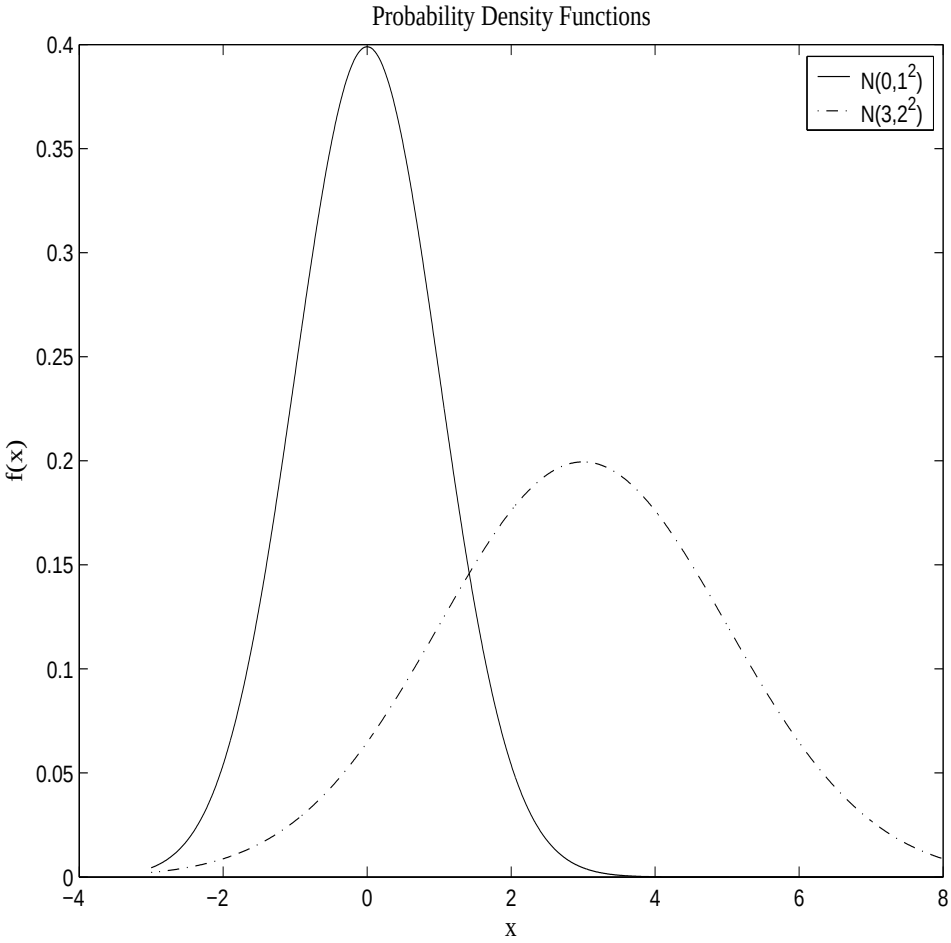

Figure 33. Case 7: Different locations, scales but same distributions: Probability density functions of $X \sim Normal(0, 1^2)$ and $Y \sim Normal(3, 2^2)$.

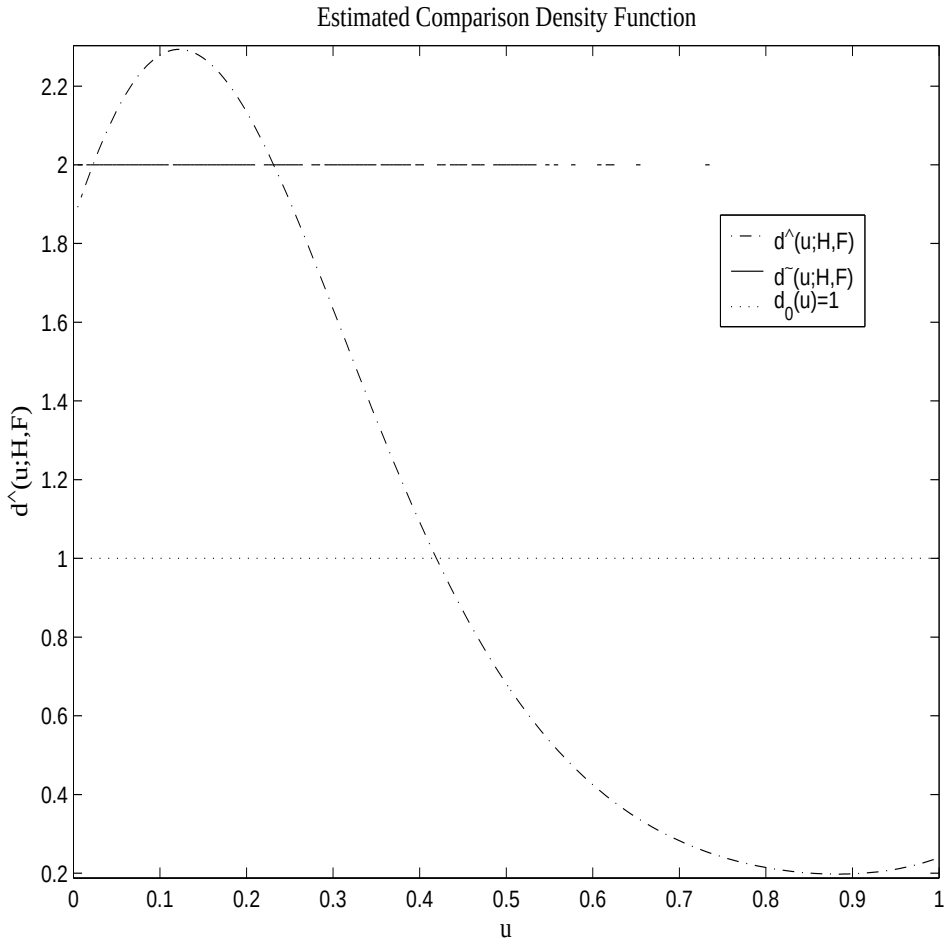

Figure 34. Case 7: Different locations, scales but same distributions: $d^{\wedge}(u; H, F)$: Estimated comparison density function with $X \sim Normal(0, 1^2)$ and $Y \sim Normal(3, 2^2)$. 1st and 3rd order score functions were selected($\mathcal{C} = \{1, 3\}$).

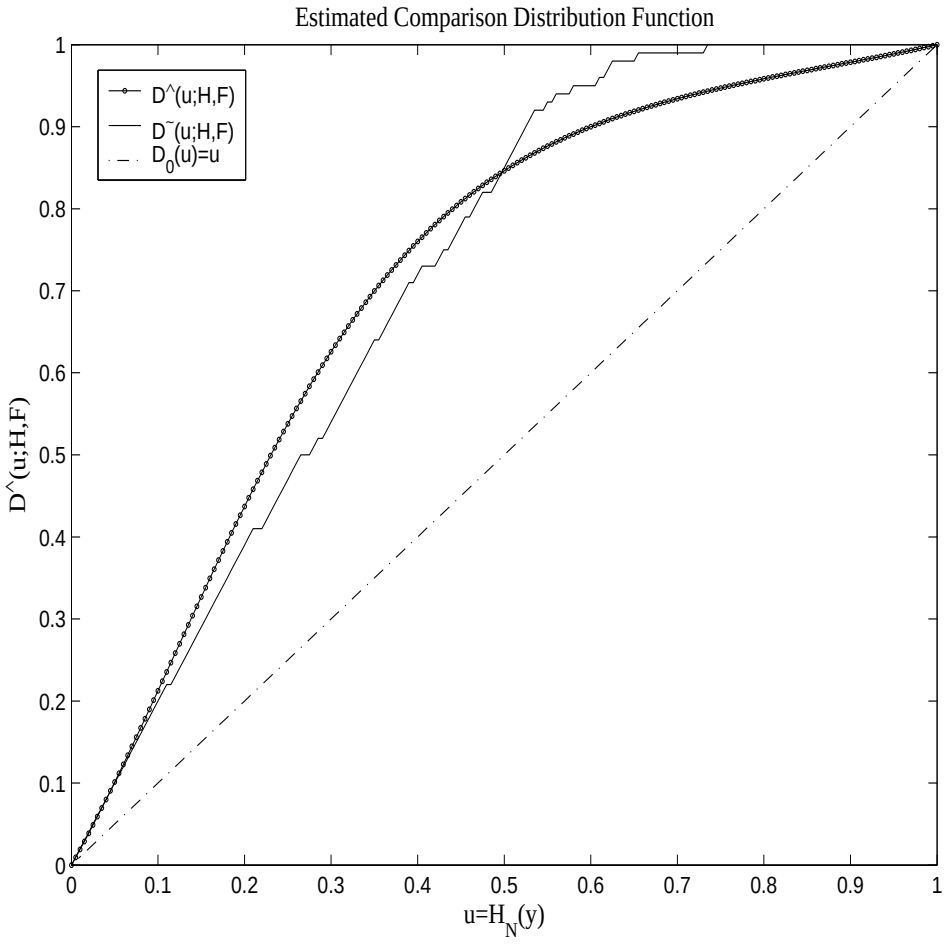

Figure 35. Case 7: Different locations, scales but same distributions: $D^{\wedge}(u; H, F)$: Estimated comparison distribution function with $X \sim Normal(0, 1^2)$ and $Y \sim Normal(3, 2^2)$

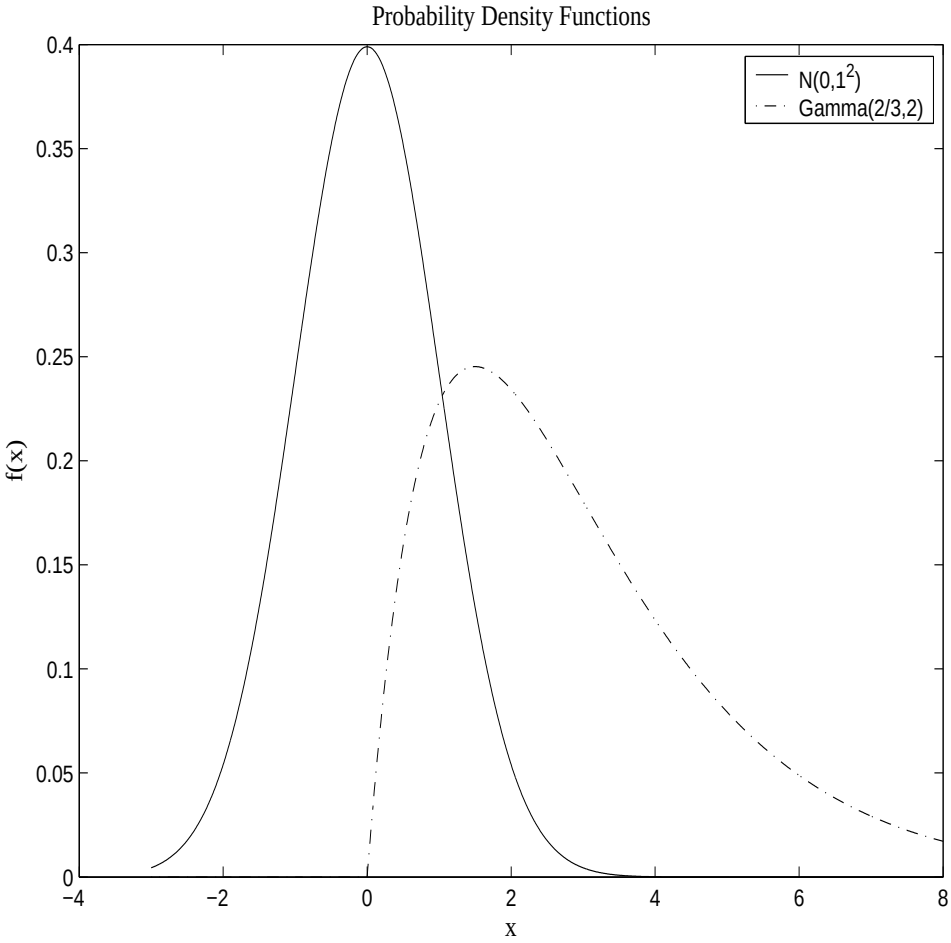

Figure 36. Case 8: Different locations, scales and distributions: Probability density functions of $X \sim Normal(0, 1^2)$ and $Y \sim Gamma(2/3, 2)$.

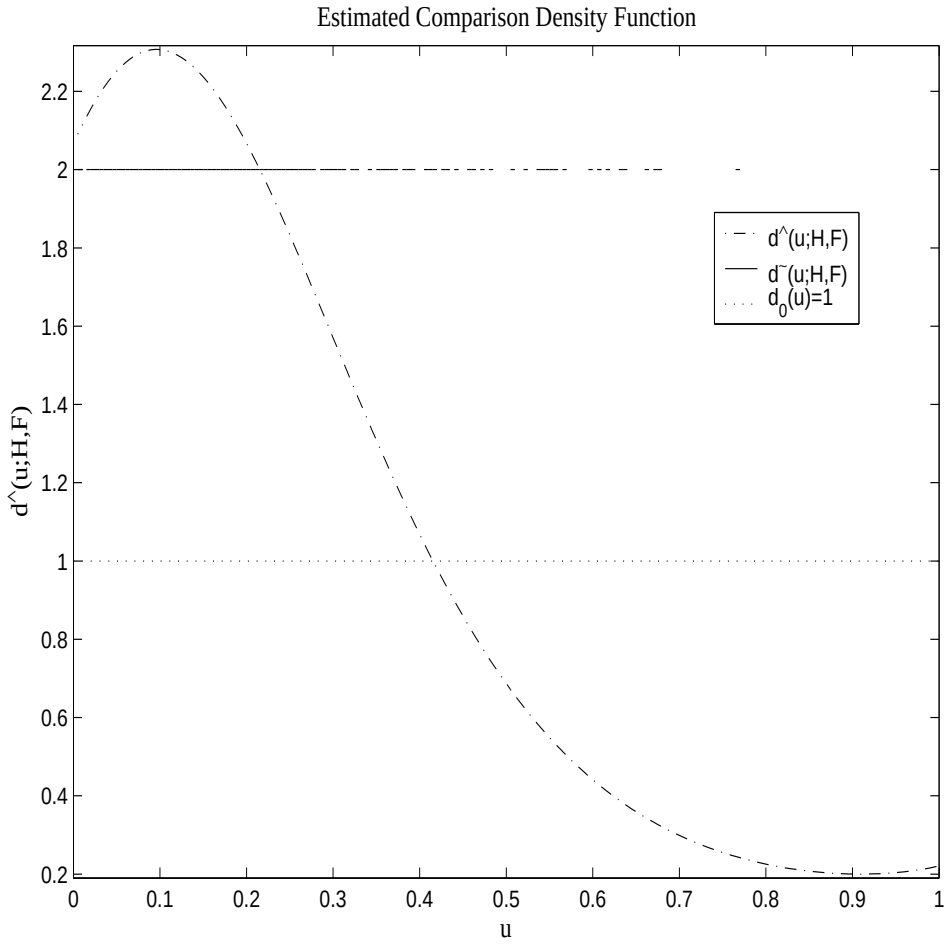

Figure 37. Case 8: Different locations, scales and distributions: $d\hat{}(u; H, F)$: Estimated comparison density function with $X \sim Normal(0, 1^2)$ and $Y \sim Gamma(2/3, 2)$. 1st and 3rd order score functions were selected($\mathcal{C} = \{1, 3\}$).

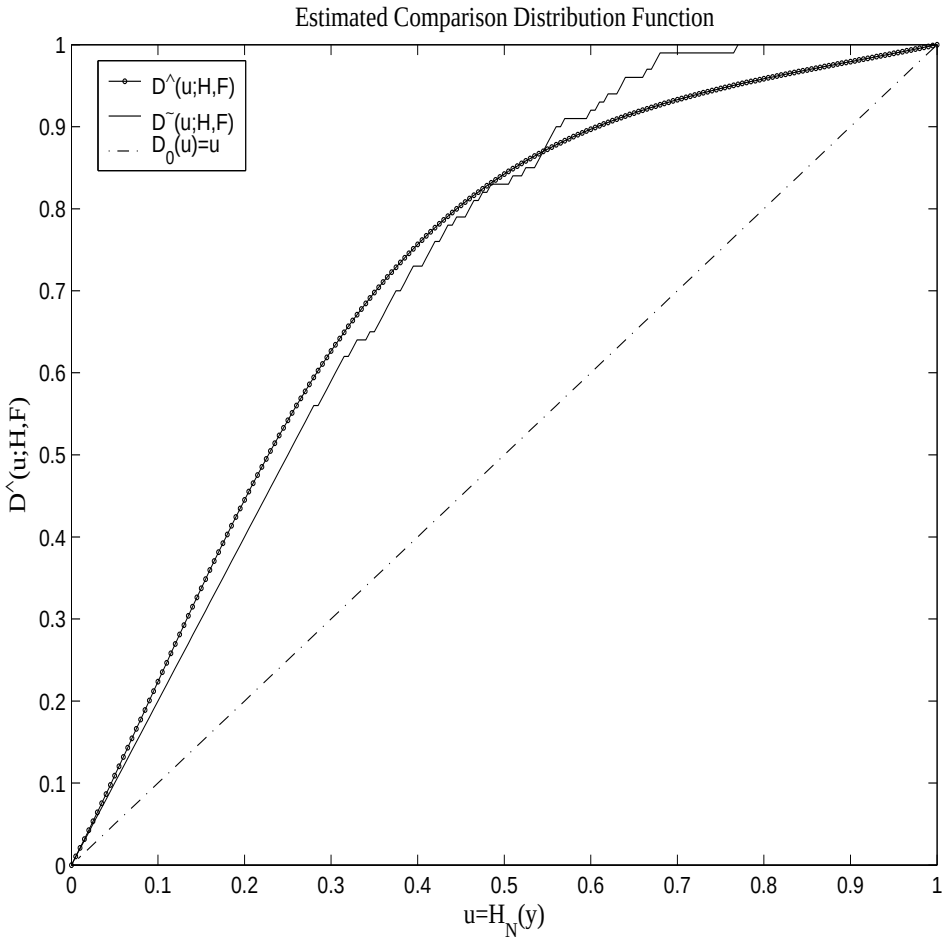

Figure 38. Case 8: Different locations, scales and distributions: $D^{\wedge}(u; H, F)$: Estimated comparison distribution function with $X \sim Normal(0, 5)$ and $Y \sim Gamma(2/3, 2)$

CHAPTER VI

CONCLUSION

6.1. Concluding Remarks

This study has aimed to discuss two-sample problem and expand the traditional two-sample data analysis. A goal of this work is to propose a two-sample data analysis procedure which is more graphical and interactive. Also, this work has sought to find a mode of analysis which provides a deeper understanding of the relation of the two populations under study.

Our exponential model approach has several desirable features that a procedure should have. It was desired to make almost no assumptions about the distribution functions of the two populations. Also it was desired to estimate the relation of distribution functions of the two populations when H_0 is rejected. In forming exponential model, it was desired to avoid doing significance test to select the significant components. Finally, it was desired to have a smooth comparison distribution function.

While reviewing existing methodologies, it was seen that the comparison density is an important object related to several of these goals. The comparison density can be used as a way of testing the homogeneity of two distribution(Under H_0, the comparison density is uniform.) and as a likelihood ratio of the density of the first sample to that of the pooled sample. Thus estimation of the comparison density is useful in the sense that it can be tested for uniformity and it serves as an estimate of the relation of the densities of two samples.

6.2. Problems for Future Study

This research has a main concern in univariate two-sample data analysis. Our future work will include

- ROC curve estimation in univariate two-sample case,
- Bivariate two-sample data analysis procedure development.

The unpooled comparison distribution function is an ordinal dominance curve(ODC) used in the evaluation of the performance of medical tests for separating two groups. And there is a relationship between ODC and receiver operating characteristic curve(ROC) such as $ROC(u) = 1 - ODC(1 - u)$. There have been not many researches related with estimation of ROC curve. Zou et al. (1997) and Lloyd (1998) proposed a smooth kernel estimator of ROC curve. And Lloyd (2002) presented a method of computing the maximum likelihood estimator of ROC curve assuming convexity. We will examine the estimation of ROC curve using our exponential model approach.

As a natural extention of univariate two-sample problem, we will examine bivariate two-sample problem through exponential model approach. Parzen (2004) gave a brief sketch of the concepts of bivariate comparison density, called dependence density , score functions and components. We will study each concept in detail and develop a bivariate two-sample data analysis procedure.

REFERENCES

Alexander, W. P. (1989). *Boundary Kernel Estimation of the Two Sample Comparison Density Function*. Ph.D. thesis, Texas A&M University, College Station, Texas.

Borovkov, A. A. (1998). *Mathematical Statistics*. British Library Cataloging in Publication Data.

Carmichael, J.-P. (1976). *The Autoregressive Method*. Ph.D. thesis, State University of New York, Buffalo, New York.

Cencov, N. (1962). Evaluation of an unknown distribution density from observations. *Soviet Mathematics* **3**, 1559–1562.

Chernoff, H. and Savage, R. I. (1958). Asymptotic normality and efficiency of certain nonparametric test statistic. *The annals of Mathematical Statistics* **29**, 972–994.

Crain, B. R. (1973). A note on density estimation using orthogonal expansions. *Journal of the American Statistical Association* **68**, 964–965.

Crain, B. R. (1974). Estimation of distributions using orthogonal expansions. *The Annals of Statistics* **2**, 454–463.

Crain, B. R. (1976). More on estimation of distributions using orthogonal expansions. *Journal of the American Statistical Association* **71**, 741–745.

Eubank, R., LaRiccia, V. and Rosenstein, R. (1987). Test statistic derived as components of pearson's phi-squared distance measure. *Journal of the American Statistical Association* **82**, 816–825.

Giampaoli, V. and Singer, J. M. (2004). Bayes factors for comparing two restricted means: an example involving hypertense individuals. *Journal of Data Science* **2**, 399–418.

Lloyd, C. J. (1998). The use of smoothed roc curve to summarise and compare diagnostic system. *Journal of the American Statistical Association* **93**, 1356–1364.

Lloyd, C. J. (2002). Estimation of a convex roc curve. *Statistics and Probability Letters* **59**, 99–111.

Parzen, E. (1962). On estimation of a probability density function. *Annals of Mathematical Statistics* **33**, 1065–1076.

Parzen, E. (1979a). A density-quantile function perspective on robust estimation. In *Robustness in Statistics*, pages 237–258.

Parzen, E. (1979b). Nonparametric statistical data modeling. *Journal of the American Statistical Association* **74**, 105–121.

Parzen, E. (1979c). Reply to comments on nonparametric statistical data modeling. *Journal of the American Statistical Association* **74**, 129–131.

Parzen, E. (1982). Data modeling using quantile and density-quantile functions. In de Oliveira, J. T. and Epstein, B., editors, *Some Recent Advances in Statistics*, pages 23–52. Academic Press, New York.

Parzen, E. (1983). Quantiles, parametric-select density estimation, and bi-information parameter estimators. In *Computer Science and Statistics: Proceedings of the 14th Symposium on the Interface*, pages 241–245, Rensselaer Polytechnic Institute. Springer Verlag.

Parzen, E. (1989). Multi-sample functional statistical data analysis. In *Statistical Data Analysis and Inference*, pages 71–84, University of Neuchatel. Elsevier.

Parzen, E. (1990). Unification of statistical methods for discrete and continuous data. In *Computing Science and Statistics: Proceedings of the Symposium on the Interface*, pages 235–242, Michigan State University. Springer Verlag.

Parzen, E. (1993). Change PP plot and continuous sample quantile function. *Communications in Statistics, Part A – Theory and Methods* **22**, 3287–3304.

Parzen, E. (1994). From comparison density to two sample analysis. In *The Frontiers of Statistical Modeling: An Informational Approach*, pages 39–56, University of Tennessee. Kluwers.

Parzen, E. (1997). Data mining, statistical methods mining, and history of statistics. In *Computing Science and Statistics. Mining and Modeling Massive Data Sets in Science, Engineering, and Business with a Subtheme in Environmental Statistics. Proceedings of the 29th Symposium on the Interface*, pages 365–374, Houston. Springer Verlag.

Parzen, E. (1998). Statistical methods mining, two sample data analysis, comparison distributions, and quantile limit theorems. In Szyszkowicz, B., editor, *Asymptotic Methods in Probability and Statistics. A Volume in Honor of Miklos Csörgő*, pages 611–617. Elsevier, New York.

Parzen, E. (2002). All statistical methods learning. Technical Report, November, Texas A&M University, College Station.

Parzen, E. (2004). Quantile probability and statistical data modeling. *Statistical Science* **19**, 652–662.

Schwartz, S. (1967). Estimation of a probability density by an orthogonal series. *The Annals of Mathematical Statistics* **38**, 1262–1265.

Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization*. John Wiley & Sons.

Scott, D. W., Gotto, A. M., Cole, J. S. and Gorry, G. A. (1978). Plasma lipids as collateral risk factors in coronary artery disease – a study of 371 males with chest pain. *Journal of Chronic Diseases* **31**, 337–345.

Tarter, M. and Kronmal, R. (1970). On multivariate density estimates based on orthogonal expansions. *The Annals of Mathematical Statistics* **41**, 718–722.

Tarter, M. E. (1979). Density estimation applications for outlier detection. *Computer*

Methods and Programs in Biomedicine **10**, 55–60.

Wakefield, M. and Kohler, J. (1991). Indoor radon and childhood cancer. *The Lancet* **338**, 1538–1539.

Woodfield, T. J. (1982). Statistical modeling of bivariate data. Research Report 23, Department of Statistics, Texas A&M University, College Station, Texas.

Zou, K., Hall, W. and Shapiro, D. (1997). Smooth non-parametric receiver operating characteristic (roc) curves for continuous diagnostic tests. *Statistical Medicine* **16**, 2143–2156.

APPENDIX A

PROOFS

• **Some properties of Comparison Density Function**

From the definition of the pooled comparison distribution function,

$$\begin{aligned} H(y) &= \lambda F(y) + (1-\lambda)G(y) \\ \Rightarrow H(Q_H(u)) &= \lambda F(Q_H(u)) + (1-\lambda)G(Q_H(u)) \end{aligned} \tag{A.1}$$

After differentiating,

$$\begin{aligned} h(Q_H(y)) &= \lambda f(Q_H(u)) + (1-\lambda)g(Q_H(u)) \\ \Rightarrow \frac{h(Q_H(u))}{f(Q_H(u))} &= \lambda + (1-\lambda)\frac{g(Q_H(u))}{f(Q_H(u))} \\ \Rightarrow \frac{1}{d(u)} &= \lambda + (1-\lambda)\frac{g(Q_H(u))}{f(Q_H(u))} \\ \Rightarrow d(u) &= \frac{1}{\lambda + (1-\lambda)\frac{g(Q_H(u))}{f(Q_H(u))}} \\ &= \frac{f(Q_H(u))}{\lambda + (1-\lambda)g(G_H(u))} \end{aligned} \tag{A.2}$$

Then from A.2, we conclude that $d(u) \to 0$ if $f \to 0$ and $d(u) \to 1/\lambda$ if $g \to 0$. Also,

$$\max{}_u\, d(u) = \max{}_u \left[\frac{1}{\lambda + (1-\lambda)\frac{g\left(Q_H(u)\right)}{f\left(Q_H(u)\right)}} \right] = \frac{1}{\lambda} \tag{A.3}$$

when $g\big(Q_H(u)\big)/f\big(Q_H(u)\big)$ has minimum value 0. Since $d(u)$ is a density function, $d(u) \geq 0$.

• **Mean and Variance of mid-distribution transform**

For the mid-distribution transform $W = F^{mid}(Y)$ defined in section 3.2.1,

$$\begin{aligned} \mu_{mid} &= E\big(W\big) = 0.5 \\ \sigma^2_{mid} &= VAR\big(W\big) = [1 - E\big(p^2(Y)\big)]/12. \end{aligned} \tag{A.4}$$

Proof : Let Y have values y_j with probability $p_j (j = 1, \cdots, n)$, $u_j = p_1 + \cdots + p_j$, $u_j^{mid} = u_j - .5p_j$, $u_0 = 0$ and $u_1 = 1$. First, verify the following equations.

$$\begin{aligned} \frac{(u_j^2 - u_{j-1}^2)}{2} &= p_j u_j^{mid}, \\ \frac{(u_j^3 - u_{j-1}^3)}{3} &= \frac{p_j(u_j^2 + u_j u_{j-1} + u_{j-1}^2}{3} = p_j \|u_j^{mid}\|^2 + \frac{p_j^3}{12}. \end{aligned} \tag{A.5}$$

(A.6)

$$\begin{aligned} \frac{u_j^2 - u_{j-1}^2}{2} &= \frac{(u_j + u_{j+1})(u_j - u_{j+1})}{2} \\ &= \frac{\big(2(p_1 + \cdots + p_{j-1}) + p_j\big)p_j}{2} \\ &= u_{j-1}p_j + .5p_j^2 \\ &= u_j p_j + .5p_j^2 - p_j^2 \\ &= u_j p_j - .5p_j^2 \\ &= (u_j - .5p_j)p_j \\ &= u_j^{mid} p_j. \end{aligned} \tag{A.7}$$

$$\begin{aligned} \frac{u_j^3 - u_{j-1}^3}{3} &= \frac{p_j(u_j^2 + u_j u_{j-1} + u_{j-1}^2)}{3} \\ &= \frac{p_j\big((2u_j - p_j)^2 - u_j(u_j - p_j)\big)}{3} \\ &= \frac{p_j(2u_j - p_j)^2 - p_j(u_j - p_j/2)^2 + p_j^3/4}{3} \\ &= p_j \|u_j^{mid}\|^2 + \frac{p_j^3}{12}. \end{aligned} \tag{A.8}$$

By using equations A.6,

$$\begin{aligned} E\big(W\big) &= \sum_{j=1}^{n} p(y_j) F^{mid}(y_j) \\ &= \sum_{j=1}^{n} \frac{u_j^2 - u_{j-1}^2}{2} \\ &= \frac{1}{2} Var(U) - \frac{u_0}{2} + \frac{u_1}{2} - \frac{1}{2} Var(U) \\ &= .5 \end{aligned} \tag{A.9}$$

$$\begin{aligned} VAR\big(W\big) &= \sum_{j=1}^{n} p_j \big(F^{mid}(y_j) - .5\big)^2 \\ &= \sum_{j=1}^{n} p_j \|u_j^{mid}\|^2 - \frac{1}{2} + \frac{1}{4} \\ &= \sum_{j=1}^{n} \frac{u_j^3 - u_{j-1}^3}{3} - \sum_{j=1}^{n} \frac{p_j^3}{12} - \frac{1}{4} \\ &= \frac{1}{3} \sum_{j=1}^{n} u_j^3 - \frac{1}{3} \sum_{j=1}^{n} u_j^3 - \frac{u_0}{3} + \frac{u_n}{3} - \frac{E\big(p(y)^2\big)}{12} - \frac{1}{4} \\ &= \frac{1}{3} - \frac{1}{4} - \frac{E\big(p(y)^2\big)}{12} \\ &= \frac{1}{12} \left[1 - E\big(p(y)^2\big)\right]. \end{aligned} \tag{A.10}$$

• Relationship between θ_1 and Wilcoxon's rank sum statistic

To compute Wilcoxon's rank-sum test statistic, combine two samples into a single ordered sample and then assign ranks to sample values. Let $R_1(i)$ and $R_2(j)$ denote ranks assigned to each sample. Wilcoxon test statistic is defined as follows;

$$T_k = \frac{W_k - E(W_k)}{\sqrt{Var(W_k)}} \tag{A.11}$$

where $k = 1, 2$ and n_k is number of observations in kth population(followed by previous section, $n_1 = m$ and $n_2 = n$), and $W_k = \sum_{i=1}^{n_k} R_k(i)$, $E(W_k) = n_k(N+1)/2$ and

$Var(W_k) = mn(N+1)/12$. From the equation(A.11),

$$\begin{aligned} T_1 &= \frac{W_1 - \frac{m(N+1)}{2}}{\sqrt{\frac{mn(N+1)}{12}}} \\ &= \left[\frac{12(N-1)\lambda_N}{(1-\lambda_N)}\right]^{.5}(R_1^- - .5) \\ &= \left[\frac{12(N-1)\lambda_N}{(1-\lambda_N)}\sigma^2_{mid}\right]^{.5}\theta_1 \end{aligned} \tag{A.12}$$

where $\lambda_N = m/N$ and defining

$$\begin{aligned} R_1^- &= (1/m)\sum_{t=1}^{m}(R_1(t) - .5)/N \\ &= \frac{W_1}{mN} - \frac{1}{2N}. \end{aligned} \tag{A.13}$$

VITA

Sujung Choi was born in Pusan, on February 4, 1973. She is a daughter of Seung-Nam Choi and Shinja Lee. She graduated from the Yonsei University, Seoul with a Bachelor of Arts degree in Statistics in 1996 and with a Master of Arts degree in Statistics in 1998. Her permanent address is Evergreen Oscarville 102-1701, Yang-San Dong, Osan City, Kyoungki Province, Korea.